A REFERENCE TEST APPROACH FOR MODERATING INTERNAL ASSESSMENT AT THE UPPER SECONDARY SCHOOL LEVEL

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Master of Arts in Education

in the

University of Canterbury

by

Glenn Stephen Chamberlain

University of Canterbury

1988

To my family ....

who put up with my presence

a little longer than expected!

In some countries it is accepted that the teacher is likely to know more about [their] pupils than an external examiner, and that [they] can provide more information about them than a necessarily short examination can hope to do.

What [they] cannot do is to be sure that [they are] accurately assessing the standards of [their] own pupils in relation to those of other pupils in other schools; this requires either positive, widely-informed and responsible moderation, or an external examination.

Schools Council, 1964.

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank Professor Warwick Elley for his dedication, guidance and encouragement as supervisor of this project. His thoughtful comments and perceptive criticisms during all stages of the project, have been of immeasurable value.

I would also like to express my appreciation to the following people who gave of their time and effort so freely: Dr David Watkins who acted as secondary supervisor for the project; Mrs Val Elley and Brian Keeling who helped with the administration of the tests; Hayati Abdullah who very kindly conducted a re-mark of the essays; Drs Bill Rosenberg and Alan Thompson (Computer Services Centre) who patiently debugged some of my programming efforts so the computer would do what it was meant to do; and Nick Fitzgerald who not only assisted with the administration of the tests and numerous other tasks, but was always there to bounce ideas off.

Finally, I am particularly grateful to the five teachers - Jim Bowden, Murray Depree, Marion Hobbs, Rodger Murfitt and Vern Tie - who acted as liaison for their respective schools, assuming responsibility for arranging suitable classes, testing sessions, supervisors etc. Last, but certainly not least, the project would not have proceeded far without the co-operation and whole-hearted effort of those 400 or so Canterbury fifth formers in 1986.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

TABLE PAGE

LIST OF FIGURES

ABSTRACT

The main purpose of this study was to investigate the suitability of reference tests for moderating internally assessed national qualifications at the upper secondary school level. In a secondary analysis, the relative merits of two alternative item formats, the open-ended and cloze, were compared with the multiple-choice item, which has been traditionally utilized in reference tests of this nature.

A series of short reference tests, based on the underlying construct of developed abilities, were constructed in four core subject areas (i.e. English, mathematics, science and social studies), along with an additional test of scholastic aptitude. The English, science and social studies tests consisted of a vocabulary and reading comprehension component; while the mathematics test had a more traditional content, relating to the measure of general concepts. An essay test was added to the English test analyses. Multiple forms of the developed abilities tests were developed as separate multiple-choice and open-ended/cloze formats, and to enable a multiple matrix sampling technique to be employed.

The validity of the reference tests was evaluated by using the performance of Christchurch fifth formers on the tests to predict their corresponding School Certificate Examination class parameters (i.e. mean and standard deviation). These analyses were based on a sample of 18 classes, across four state, co-educational high

schools; covering a wide range of ability levels. A series of multiple regression analyses were conducted to provide optimal predictions of the respective class parameters.

It was found that each of the subject-based reference tests predicted class ability levels (i.e. means) on the corresponding School Certificate Examinations with a very high degree of sensitivity. The multiple-R's generated were 0.97 for mathematics, 0.90 for English, 0.89 for science and 0.80 for social studies. The predictions of the spread of ability for each class (i.e. standard deviation) were found to be more difficult, although the results were still sensitive enough for moderation purposes.

The addition of the mathematics or scholastic aptitude test to the subject-based reference tests improved the multiple-R's on both parameters.

The comparison of item types revealed no significant difference in the prediction of class means. However, the open-ended/cloze format failed to predict class standard deviations at a statistically significant level.

The findings were discussed with reference to those from earlier studies, to policy implications, to application at a practical level and to the urgent need for further research.

CHAPTER I

INTRODUCTION

> The examination dictates the curriculum
> and cannot do otherwise; it confines experiment,
> limits free choice of subjects, hampers treatment
> of subjects, encourages wrong values in the classroom.
>
> Ministry of Education, London, 1960.

GENERAL BACKGROUND ISSUES

There is one important feature of our education system which has remained largely unchanged over the past 40 years and until very recently, has provided us with only limited encouragement. The problem area concerns the approach taken in our formal assessment of pupils at the upper secondary school level.

In the period since the Second World War and up until 1986, pupils and teachers have been forced to cope with a rigid three-year assessment structure, comprising four external (or national) examinations. School Certificate, a Department of Education award, is normally attempted at the end of Form 5 after three years of secondary schooling. Candidates may sit up to six subjects, with the majority attempting four or five. University Entrance - although recently abolished - has been the major Form 6 award, and was administered and controlled by the Universities Entrance Board. Successful candidates then had the choice of either going straight on to university or preparing

further by completing a Form 7 course. At this level two further university-based qualifications are offered, the Bursary Examination which is attempted by nearly all Form 7 pupils, and the more demanding Scholarship Examination, generally attempted only by the more able students.

This system of examination-oriented assessment has remained virtually untouched over the years except for minor modifications, such as the introduction of single-subject passes for the School Certificate Examination in 1968. Even the arrival of the new internally assessed Sixth Form Certificate in 1968 - awarded to all pupils on the completion of any Form 6 course - had until 1986 been largely overshadowed and restricted by the more prestigious University Entrance award.

During this period, our examination system has resisted all progressive change, which has been in direct contrast to other aspects of New Zealand life. Rapid social change has made increasingly greater demands on our education system for a multitude of reasons, for example the technological and computer revolution. No longer is it simply the case that pupils not bound for university could afford to leave school with just a pair of "unskilled" hands to impress employers with. The pool of manual positions available on the job-market has gradually decreased, while the corresponding increase for skilled-manual and non-manual positions has placed new demands on the secondary schools, poly-technics, universities and other similar training institutions.

In the secondary schools between 1945 and 1965 the majority of pupils left after Form 5 to obtain a job. However, by 1985 this majority was now staying on for an additional Form 6 year to increase their likelihood of obtaining a job in what is now a highly competitive labour-market. The higher retention rate at the Form 6 level suggests this group can no longer be regarded as an "academic elite" heading for university. In fact, the most recent figures available show that only 12.5 percent of pupils leaving secondary school intend going on to university (Department of Education, 1984), and even a smaller minority of these would be doing so directly from Form 6.

The consequence of this trend has been an increasing demand for new and broader courses providing relevant knowledge and skills, which are more readily sought after by employers. However, up until 1986, the need to expand the range of courses offered in Form 6 in the form of Sixth Form Certificate qualifications was being constrained by the continued presence of the more prestigious, but totally academically-oriented University Entrance Examination. The influence of the university-based examinations and the School Certificate Examination on the range of subject choice was becoming too great. The second Ministerial Review on Curriculum and Assessment, 1986 (hereinafter to be known as the Ross Report) similarly observed that,

> ... the demands of examinations such as School Certificate and those administered by the Universities Entrance Board seriously restrict the opportunity for students in the senior secondary school to take a sufficiently wide range of subjects to lead to the broad and general education envisaged. (p 45)

The examination method of assessment has been firmly entrenched in our secondary schools, partly through the influence of the Universities Entrance Board's requirements at the Form 6 and 7 level, and also partly as a "throwback" to the community's perception of accountability, maintenance of minimum standards and equality of educational opportunity. Over the years, the traditional examination has come to be accepted by the public as the only reliable and valid method of evaluation, and the mere thought of making changes can often evoke heated and emotional responses as indicated by the content of editorials and letters in the daily newspapers.

However, the continued use of external examinations in their present form must be questioned in light of their constraints on an evolving society which is demanding a longer and broader education. Their major disadvantage is the "backwash" effect in the classroom, where only the examinable syllabus material is taught. This results in a narrow focus by the teacher and pupils, directed solely at passing the examination. A further consequence is the amount of influence it has on what is taught in earlier courses.

Other disadvantages include the narrow focus on writing skills at the expense of other important skills (e.g. oral, practical, interest and attitude); the stress and chance factor associated with a one-shot-three-hour examination; the lack of detailed information about each pupil's relative strengths and

weaknesses; the weak predictive ability between external examinations and non-academic occupations; the difficulty with the development and implementation of relevant local courses; and the inconvenience for pupils wanting to plan ahead when examination results are not returned until late January.

The Post Primary Teachers' Association (PPTA) has been campaigning strongly since the early 1970's for the replacement of external examinations with internal assessment at the Form 5 and 6 level. Numerous articles have appeared in the PPTA Journal (e.g. Bray, 1971; Bristow, 1971; Capper, 1987; Elley, 1976 and 1985; Hall, McMurray and Capper, 1985; Hogan, 1976; Hughes and Keeling, 1976a and 1976b; McGaw, 1974 and 1976) exploring the various issues with the aim of stimulating discussion among its members.

There are two major reasons which explain the sluggish progress made in reforming New Zealand's national examination system. Firstly, the effects of a lengthy period in power by the National Government (1975-1984) produced little or no educational reform despite the many forward-thinking recommendations produced in the Educational Development Conference (EDC) Reports (1974), the McComb's Report (1976), and the general trend away from external examinations to partial or complete internal assessment in other Western nations (e.g. Australia, Canada, Sweden, UK and USA). Indeed many commentators felt education was regressing. In particular, the effects of a very conservative Minister of Education, Mr Merv. Wellington (1979-1984), resulted in only bitter

arguments with the PPTA and other reforming groups. His view of assessment was staunchly in favour of full external examinations at each of the last three form levels of the secondary school.

However, the election of a new Labour Government in 1984 with a greater expressed concern for the needs of pupils and teachers, eased many of the earlier tensions and allowed for the implementation of some reforms. These included the removal of the University Entrance Examination from Form 6, thus allowing the internally assessed Sixth Form Certificate to stand alone as the major award at that level; the removal of the "pass-fail" concept from the School Certificate results; and the inception of Ministerial Reviews of the curriculum and the assessment system to produce a set of policy proposals.

The second reason relates to a serious lack of research focussing on the many assessment issues in need of investigation. Unfortunately, some of the recommendations in the Ross Report (1986), for example the intended move to achievement-based testing, are likely to remain as proposals until much needed research has been completed to ensure a smooth transition from external examinations to internal assessment at the senior secondary school level. One of these recommendations, technical in nature, involves the development of suitable moderating procedures through which teachers can compare and maintain the standards of achievement attained by their pupils, relative to pupils of other classes, schools, year and subject groups. Among its list of recommendations for assessment, the Ross Report (1986, p 16) stated that,

> ...urgency be given to the setting up of trials to investigate alternative moderation procedures for sixth form subjects.

As an indication of the lack of progress made, a similar recommendation was reported over a decade earlier:

> Research should be initiated immediately to investigate the feasibility of using standardized tests and other procedures to moderate teachers' assessments in different schools. (EDC, 1974, p 13)

The current method of moderating Sixth Form Certificate is based on the marks gained by the candidates of each individual school on the previous year's School Certificate Examination. However, the non-allowance for maturation effects in pupils and the proposed abolition of the School Certificate Examination from Form 5 means that an alternative method of moderation at the Form 6 level must be investigated. Survey results of teachers and employers (EDC, 1974 and Elley, in progress) have clearly indicated a preference for some form of moderation, while it has been argued at some length by Elley and Livingstone (1972) that for internal assessment to operate successfully the development of a sound moderation scheme is essential. In addition, it would seem most unfair when competition for jobs and further education is at its highest, that a suitable method of maintaining comparable standards between schools is not implemented.

With this in mind, it is unfortunate that the sum total of major, relevant empirical studies within New Zealand can be summarized thus:

> Two theses by Alison Gilmore (Otago, 1979) and Murray Hulbert (Waikato, 1978) have explored some of the problems of these [moderation] approaches, but much more research is required. (Elley, 1985, p 14)

In view of the lack of knowledge about suitable moderating procedures and its related importance for policy changes, the current study was initiated to investigate an alternative approach to moderation.

STATEMENT OF THE RESEARCH PROBLEM

The Ross Report suggested several possible approaches to moderation including the use of reference tests or achievement-based assessment (p 64). The latter method has received the greatest amount of attention, especially from the PPTA, as the assessment option for the future. However, there has been little significant research in this area and research offerings from overseas have also been limited. In addition, the research and retraining of teachers required to implement this approach will be much more than that required for the introduction of reference testing. As a short to medium term solution to the problem of moderation, the use of reference tests appears to be the most feasible option.

This method is already used to moderate mathematics and science at School Certificate level. The procedure involves all pupils sitting a common test at some point during the Form 5 year to produce a fair range of grades for allocation in a particular subject or school. There has been little research in relation to this policy proposal and most of it has focussed on the

use of rererence tests with an exclusively multiple choice format. Some subject teachers, English especially, are becoming increasingly less convinced as to the validity of employing just a single objective item type. Thus, it would seem desirable to assess the accuracy of alternative item formats for use in moderating tests.

Thus, the specific focus of the current study can be viewed as a two-fold investigation:

(1) To develop and validate reference tests in four core subject areas - English, mathematics, science and social studies. With the addition of a general ability test, the aim is to assess the sensitivity, both singly and in combination, with which the class parameters (mean and standard deviation) of School Certificate Examinations may be predicted from pupil performance on the reference tests; and

(2) To assess the suitability of three different question formats - open-ended, multiple-choice and the cloze formats. An attempt was made to analyze which question types, singly and in combination, best meet the demands of the respective subject areas.

CHAPTER II

REVIEW OF THE LITERATURE AND DEVELOPMENT OF THE RESEARCH AIMS

WHAT IS MODERATION?

As stated in Chapter I, moderation is concerned with maintaining standards of educational achievement between different groups, such as classes, schools or subjects. Rowlands (1974, p 85) described the process quite simply as,

> ... a method of maintaining comparability of standards between schools [, classes, subjects or year groups] usually without prescribing detailed syllabus content or requiring students to sit for an external examination.

However, the assumption of maintaining comparable standards is itself, an important consideration.

IS MODERATION NECESSARY?

Traditionally, New Zealand has been regarded as an isolated and democratic haven, free of most social barriers, like class and race, and in many ways our use of the term "Godzone" has not been without foundation. Similarly, in our schools there is a traditional belief in equality of opportunity and openness, where the only barrier to academic success was one's-own-self. Throughout the high schools and colleges pupils were learning the same courses, teachers taught the same syllabi and pupils nervously awaited the same examinations on the same day.

This kind of standardization gave the appearance of an objective and fair form of assessment and also produced an effective means of maintaining comparable educational standards between schools, subjects and year groups.

The element of fairness and equal opportunity - that an examination mark of 85 for School Certificate English was the same whether your child attended a private single sex college or a state co-educational area school - has remained a strong feature of our education system. While there is a growing realisation today that social barriers do exist in New Zealand, and that mere opportunity is not always sufficient to achieve well at school, there still remains a strong fundamental public belief in the notions of maintaining standards and equality of opportunity. They have appeared repeatedly as arguments against the introduction of internal assessment, in the form of public protests in the letters section of our daily newspapers, and even teachers have expressed similar fears in a national survey this year:

> It is important to set a nationwide standard, or else we may get the 'good' school 'bad' school problems as seen in North America.
>
> ... some form of national moderation will be necessary to prevent differing values/standards throughout New Zealand.
>
> I feel the community will require the system to have some form of moderation to satisfy themselves that all awards by all schools are consistent and hence of equal value. (Elley, in progress)

Unlike schools in the United States, for example, where both norm- and criterion-referenced testing takes place without

any moderation and has resulted in the setting up of a complex system of standardized tests by the College Entrance Examination Board, it may be too much to expect all sectors of our community to accept such a large change at all three levels of senior school assessment. In addition, it would seem most astute to take advantage of our relative smallness by developing a feasible method of national moderation.

In accepting the assumption of educational comparability for achievement levels and Government policy to introduce internal assessment at both the Form 5 and 6 levels along with a suitable form of moderation at the latter, then there is an urgent need to look at an alternative method of moderation. In relation to this issue the Ross Report (1986, p 63) commented that:

> In the absence of external examinations it would be necessary to institute a system of moderation to ensure that the quality of assessment is uniform throughout the schools of New Zealand.

Elley and Livingstone (1972), in their classic review of the subject, the EDC Report on Directions for Educational Development (1974) and the McComb's Report (1976) have also argued the fundamental requirement of moderation in our assessment of senior secondary school pupils. Interestingly, the recommendations have changed little in more than a decade, a sure indication of the slow response to the problem:

> To permit the transition to full internal assessment to occur speedily, it is necessary to devise moderating procedures to ensure that levels of student performance are maintained under the new regime and that the standards set by different schools are comparable.
>
> (EDC, 1974, p 33)

HOW WELL CAN TEACHERS MAKE MODERATION JUDGEMENTS?

While a few secondary teachers have gained some experience with internal assessment "experiments" (e.g. the Canterbury Mathematics Scheme) it is likely that for the large majority of teachers the prospect of coping with total internal assessment remains an unknown quantity, since in the past, all major senior school credentials have been moderated by a compulsory external examination. Even the internally assessed Sixth From Certificate has continued to be moderated by the previous year's School Certificate Examination results.

There have been a number of studies, both overseas and local, which have investigated this aspect by focussing on the reliability and validity of teachers' assessments of their pupils' performances (McClelland, 1949; Yates and Pidgeon, 1957; Petch, 1964; Sowell, 1970; NZ Department of Education, 1971; Hulbert, 1978; Gilmore, 1979; Murphy, 1979 and 1981; McCausland, 1981, Wilson, 1982)

An Educational Development Conference (1974) national survey - although rather dated now - included one question relevant to the issue which identified that...

> ... 72 percent of teachers felt confident that their assessments were at least as valid as those of the external examinations.

What available evidence is there to test the validity of this claim? A recent review of related studies by two local researchers (McCausland and Hall, 1985) identified a number of

important findings in their analyses of teachers' predictions of School Certificate Examination results. Included among these were: (1) that the variability in the correlations obtained showed a significant difference between subjects (highest for English and lowest for mathematics and French); and (2) that teacher's estimates displayed a marked variation in standards relative to the achievement of their individual class groups and that this trend varied from one subject to the next (the least amount of variation occurred for English and the highest for mathematics and French). Despite several encouraging results - for example, median correlations ranging from 0.72 to 0.89 (lowest for English and highest for mathematics and French) and a relatively high degree of agreement (83^{+}_{-} percent) between teachers and examiners in classifying pupils' School Certificate total scores as pass or fail - the authors still felt bound to conclude that,

> ... for those readers with particular responsibilities in the field of examining and testing, for example, given the increasing involvement of teachers in internal assessment, the [first two findings] reaffirm dramatically the need for scaling or monitoring if the assessments of teachers are to be linked to a common standard. (Ibid. p 91)

McCausland and Hall's findings suggest the original claim of confidence by 72 percent of teachers may be reasonably accurate, but the size of the discrepancy among the rest of the teachers leaves one with little confidence about the reliability of such predictions. Likewise, but in reference to Form 6 assessments, Elley and Livingstone (1972), reported a significant discrepancy in the accrediting procedure for University Entrance,

to the extent, "that over 60 percent of schools" are likely to deviate from the national norm by a greater margin than the error measurement would account for. In relation to the theme of educational comparability the authors' concluded that:

> Even with a back-up examination to serve as a check on faulty estimates, it is clear that schools have difficulty in setting consistent standards when making internal accrediting decisions. Without such an aid, their problems would be even greater.
>
> (Ibid. pp 61-2)

What can overseas research reveal about the need for moderation? As mentioned earlier, a high school leavers' assessment system without moderation, such as that in the United States, can cause a similar backwash effect in schools as one based on external examinations. The development of a national system of college entrance aptitude testing to assist in university entrance decisions, has resulted in the teachers and schools teaching towards the entrance examinations in the same way that teachers taught to external examinations in New Zealand (Elley, 1976, p 29). This "backwash" effect into the school curriculum has certainly been prevalent in Grades 11 and 12 and to a lesser degree in Grade 10 in the USA (Keepes and Keepes, 1974), but the parallel caused by a similar effect from our University Entrance and School Certificate Examinations is striking.

The strongest evidence so far, for moderation, comes from Canada. Several reports by Elley (1976, 1985) have focussed on the recent shift from external examination to internal assessment.

Several of the Canadian provinces - British Columbia, Alberta and Ontario - introduced internal assessment without the aid of any system of moderation. Almost immediately, the consequences were felt as some schools "inflated" their grades while other schools maintained the status-quo. For example, in Ontario when examinations ceased in 1968, the number of pupils qualifying for university admission increased by 20 percent in just the first year of the new assessment system. This increase continued in subsequent years. In British Columbia and Alberta the trends have been very similar, but with the unfortunate consequence of having external examinations reintroduced in Grade 12 to assess 50 percent of the year's work and to act as a system of moderation. A similar fate seems likely for Ontario as well:

> The Canadian education authorities neglected to develop a proper moderating system, and are now going through the painful process of reintroducing a form of examinations in Grade 12... the backward swing of the pendulum was in my view almost entirely due to an absence of moderation. (Elley, 1985, p 11)

On the Australian scene, the introduction of internal assessment has fared much better and is now firmly established at the Year 10 (Form 5) level in all eight states. A crucial factor in this success has been the development of suitable moderation procedures. For example, a test-based approach has been used in New South Wales and Tasmania, the former moderating in only English and mathematics after starting in over 30 subjects in the mid 1970's, while Queensland and Victoria have adopted a complex system of regional moderation, as has Western Australia, but with additional reference testing as well (Radford, 1974, p 16).

At the Year 11 and 12 level most of the states still retain an examination for the assessment of their more prestigious "Group One" (i.e. university entrance) subjects, although in most cases there is an internally assessed component ranging in weighting from 30 to 50 percent. The "Group Two" subjects (i.e. school based, non-continuing) are totally internally assessed and are moderated by either a regional approach or by using reference tests. (McGaw Report, 1984). However two states, Queensland and Australian Capital Territory (ACT), have acted in a more progressive manner by introducing internal assessment at this level also, thus removing all external examinations from their secondary schools. Initially, the Queensland educational authority continued with their regional teacher-based moderation scheme, but have now elected to use the nationally normed Australian Scholastic Aptitude Test (ASAT) in the final year. Likewise, the ACT authorities have also chosen to use ASAT as their major method of moderating educational standards in Grade 12.

The final piece of evidence comes from Sweden, where internal assessment has been in existence since the 1930's. There, the National Board of Education has used standardized (or reference) tests to act as a moderating device, "to enable the teacher to compare the performance of [their] own class with that of the total population and adjust [their] marking according to the outcome of the testing" (Marklund, 1985, p 11). The basic technique appears to have changed little over the past two decades (Henrysson, 1964) and also seems to have stood the test of time without any major technical problems.

In the light of the evidence gathered from the overseas research, it is imperative that a suitable method(s) of moderation be developed to maintain educational comparability between classes, schools, subjects and year groups as New Zealand makes the transition from external to internal assessment. As a confirmation of the overseas trends, one section of a 1987 national survey of 500 senior secondary school teachers on assessment policies and moderation revealed that,

> ... only 2 percent of teachers at the Form 5 level and only 3 percent at the Form 6 level, would prefer internal assessment without moderation.
> (Elley, in progress)

The research evidence, although somewhat thin on quantity, has never-the-less been consistent and suggests but one conclusion:

> If the New Zealand public and the tertiary institutions are to accept internal assessment, we too must have a system of moderation.
> (Elley, 1985, p 11)

TYPES OF MODERATION

Gilmore (1979) identified two broad categories by which moderation can be classified, namely; "teacher-based" and "test-based".

Teacher-based moderation involves all methods that exclude the use of any form of common testing. The critical feature with this type of moderation is that teachers' assessments of their pupils' performances undergo adjustment based on discussions with other teachers until an agreement has been reached about common

standards of achievement to which all pupil work is then related. There are various methods, or combinations of methods, which are employed for this type of teacher or consensus based moderation. Gilmore (1979, p 12) listed some of these as:

(1) teachers adhering to a predetermined standard of marking for common tasks set for all pupils;

(2) teachers remarking their own pupils' work in the light of experience gained from evidence of other teachers' standards;

(3) reassessment of pupils' work by a group of teachers to modify individual teachers' standards; and

(4) inspection of school assessment programmes by externally employed moderators.

Teacher-based moderation is currently in use throughout Australia, in some cases as a "back-up" to test-based moderation, but more importantly as a means of moderating non-academic (or Group Two) subjects which usually contain a large practical component (e.g. art, computer science, music and physical education) and/or academic subjects which have small classes (e.g. minority ethnic languages). Several studies have evaluated this approach (Bagnell and D'Cruz, 1971; Radford, 1974; and McGaw, 1984), each noting the success of teacher-based methods within the framework described above, as a valid and reliable system of moderation. The main disadvantage is that it involves a large amount of teacher time and in some rural areas a lot of travelling. Its value as a means of evaluating practical subjects has already been realized in New Zealand where Practical Art has been assessed in a similar way for a number of years.

Test-based moderation, on the other hand, involves the use of pupils' performances on a particular test, or battery of tests, as a means of indicating the range of grades to be allocated for that specific population. The various test formats and materials available to act as a moderating device can range from a traditional external examination, such as School Certificate, to a commonly used intelligence test.

Test-based moderation is currently being used in several Australian states at different year levels. For example, New South Wales and Tasmania have used this approach to standardize their Year 10 School Certificate award (Elley, 1985, p 13), while more recently, Queensland and ACT have employed ASAT to moderate Group One subjects at the Year 11-12 level for their senior school leaving and teritary entrance certificates (McGaw, 1984). It seems likely test-based approaches to moderation will dominate the assessment of Group One subjects in most other states before long. As mentioned earlier, test-based moderation has also been used very successfully in Sweden (Henrysson, 1964; Marklund, 1985), over a long period of time, and without major technical problems.

So which type of moderation, teacher-based or test-based, would be most suitable for New Zealand? The one aspect which is clear from the research findings is that no one approach will be suitable to moderate all subjects. It has already been shown with the successful Australian and Swedish systems, how the more popular academic (or Group One) subjects are moderated by test-based approaches and the more practical (or Group Two) subjects - with

fewer numbers - are generally moderated by teacher-based approaches. An example of Group One subjects being moderated by the latter approach occurred in Queensland, when their public examinations were abolished in 1970, following the recommendations of the Radford Committee. Moderation was attempted in all subjects using a teacher-based scheme which involved the state being divided into ten moderation districts, with subject area moderators (one per school) attending district meetings and where assessment programmes and samples of pupils' work were examined and discussed (Fairburn, McBryde and Rigby, 1976).

However, despite the extensive, almost complex, nature of the system it did not appear to achieve its aim. A review study found that the majority of teachers were unhappy with the moderation system, in particular they felt that it,

> ... does not achieve comparability ... the number of variables was too great to allow meaningful comparisons between schools... time constraints [and] a large number of scripts ... encouraged moderators to make superficial subjective judgements. (Fairburn, et al. 1976, p 149)

The authors went on to conclude that,

> ... as tertiary entrance and other selection procedures associate themselves more with ASAT and less with semester ratings, the worthwhile elements of internal assessment will become more evident [and] moderation will fade as a significant issue. (Ibid, p 150)

Not unexpectedly, Queensland's educational authorities were quick to utilize the development of ASAT as a more suitable and objective means of moderating the very large numbers doing Group One subjects, and to leave teacher-based moderation as a more feasible approach for Group Two subjects with their lower numbers and more practical assessments.

What then, are the preferences of New Zealand teachers with respect to potential moderation schemes? A survey conducted by the School Certificate Examination Board in 1972 revealed that teachers clearly favoured a test-based approach to moderation, administered near the end of the school year. However, the precise nature of the test format was hotly debated (SCEB, 1974).

Elley's (in progress), 1987 national survey of 500 teachers on assessment policies and moderation, provides the most recent statement on this issue. Teachers were asked to rank their first three choices to indicate personal preference for separate assessment systems in Form 6 and Form 5. The results from the survey are set out below in Table 2.1.

TABLE 2.1

Teachers' Preferences for Potential Assessment Systems at the Form 6 and Form 5 Levels

Assessment Systems:	Form 6	Form 5
(a) Internal Assessment, Moderated by Nationally Defined Criteria Levels	26.4%	25.0%
(b) Internal Assessment, Moderated by Reference Tests	20.5%	17.4%
(c) Partial Internal Assessment/External Exam	19.6%	29.4%
(d) Internal Assessment, Moderated by S.C.Results (as at present)	9.4%	N/A
(e) Internal Assessment, Moderated by Teacher Consultation Amongst Schools	9.1%	8.3%
(f) External Examinations Only	6.7%	13.9%
(g) Accrediting (similar to former U.E.)	4.5%	2.8%
(h) Internal Assessment, Without Moderation	3.1%	2.2%
(i) Other	0.7%	1.0%
Total	100.0%	100.0%

Not surprisingly, the option of internal assessment, moderated by applying nationally defined achievement levels, (i.e. criterion-referencing), has gained a lot of support among teachers through the efforts, initially of the PPTA, and now also in the

recommendations of the Ross Report (1986) to the Government. It is clear from Table 2.1, that criterion-referenced measurement is the most preferred form of assessment in Form 6 and the second most preferred in Form 5. However, the intended introduction of criterion-based assessment may be further away than many of its proponents envisage.

Criterion-referenced (C.R.) assessment is already well established in the USA and has been introduced at lower year levels in Australia and Britain. Traditionally, secondary schools have used norm-referenced (N.R.) based assessments which describe a pupil's achievement relative to others of a similar age or grade, the aim being to differentiate pupils on the basis of their relative ability and achievement. C.R. based assessment, on the other hand, is used to ascertain what pupils have learnt with respect to specified course objectives, unrelated to the achievement of others in the same age or grade group. Thorndike (1971) and Nitko (1983), described the function of C.R. tests as that of, "yielding measurements that are directly interpretable in terms of specified performance standards". For example, a pupil doing typing achieves a rate of 60 words per minute on a test, under standardized conditions. The pupil could receive an "A" grade because s/he is near the top of the class (norm-referenced). Or, the pupil may be described as achieving a rate of 60 words per minute - under specified conditions (criterion-referenced).

While the specification of criteria for practical subjects such as typing is relatively straightforward, for Group One subjects, such as English, science and social studies, the task is proving to be rather cumbersome and time consuming. An example of some of the problems encountered can be demonstrated from a Christchurch high school, Sixth Form Certificate, geography course employing an "experimental" assessment scheme based on a C.R. system.

The notes to pupils and parents inform them that assessment will be made by comparing each pupil's work against a set of "explicitly stated criteria". Furthermore, standards will be maintained by applying criteria "which describe pupil achievement at six levels", and that, "inter-school comparability" will be achieved through "teacher consensus panel meetings" where representative samples of pupils' work will be check-marked. A review of the course criteria suggests that it would be extremely difficult for teachers to maintain even a reasonable degree of comparability when one is expected to make objective assessments using key descriptor words such as, "all", "most", "some", "comprehensive", "wide range", "simple", "minimum", "in depth", "describe" and "explain". Two examples of the assessment criteria are listed below:

EXAMPLE 1: The student...

A. can recall a comprehensive range of facts in detail

B. can recall a wide range of facts in detail

EXAMPLE 2: The student...

A. is able to recognise a range of value positions

B. is aware that different value positions exist

Even within a group of highly motivated and skilled geography teachers, each teacher's interpretation and subsequent selection of a particular achievement level to correspond with the assessment of their pupils' work, will be largely a subjective evaluation. One teacher's understanding of where "some" ends and "most" begins will most likely differ from that of a colleague in the same school, but will almost certainly differ from colleagues' in the neighbouring school. What would happen if two teachers from different schools awarded identical grades for their pupils' work, yet one has provided their class with a fully worked example similar to the question under assessment? Any such attempts to maintain inter-school comparability using supposedly explicit key descriptors, such as those listed above, will be fraught with difficulties and subsequent time constraints.

A study by Archer (1984) on the introduction of C.R. assessment at Year 12 in Queensland only reinforces the complexity of this problem. Teachers, without adequate training, were forced to develop written objectives, review methods of assessment and define criteria for marking pupils' work resulting in rivalries and misunderstandings. Many teachers considered the criterion-based system was no better than norm-referenced techniques, yet

the former required a great deal more of their time, with much resentment among teachers. Many of these same sentiments have been expressed by New Zealand teachers in a recent survey (Elley, in progress). A significant proportion of those teachers who favoured C.R. assessment for moderation purposes, did so only under the assumption that adequate training, resources and time would be made available:

> Teachers would require a great deal of in-service training and support. C.R. is a good system, BUT again, no one recognises the huge time input required (i) initially to set up, and (ii) to test the level of each student. Teachers need to get back to teaching and leave assessment to outside agencies.

Not surprisingly, Queensland has now adopted ASAT instead, to moderate its Group One subjects.

Ebel (1972) and Power (1986), highlighted the same dilemma from similar perspectives. Ebel identified three major problems: (1) C.R. measurements do not inform us of a pupil's relative ability, whereas, norm-referenced measurements can be readily provided with equivalent content descriptions; (2) C.R. measurements are unsuitable for subjects where the emphasis is on knowledge and understanding (i.e. Group One academic subjects). "Knowledge does not come in discrete chunks that can be [behaviourally] defined and identified separately." (p 147); and (3) C.R. measurements are necessary only for a small number of important educational achievements. Thus, it can be inferred that achievement is more often associated with degrees of learning rather than all-or-nothing attainments.

Power's (1986) review of the Australian position focusses on the overdrawn distinction between norm and criterion - referenced assessment from three different perspectives - the educational, political and technological. In the latter, he identified several major difficulties that need to be confronted:
(1) Domain specification and curriculum design. "Dividing upper secondary courses into domains and specifying what is in them is not easy... Nor does it reveal how knowledge, skill and process in a domain are integrated in such a way as to enable us to recognize different levels of competence." (Ibid, p 270);
(2) Fuzzy domains and test specifications. In the subject areas where domain criteria become difficult and results in many "fuzzy" or grey definitions, the expectations demanded by the test constructor may be too great. Consideration should be given to a possible trade-off between the fuzziness of defining criteria and the objectivity of testing; and (3) Specifications for developing domains. Power has listed eight aspects which the development of any domain should attempt to satisfy. Included among these is the need to reflect the nature of the subject; be capable of being identified and measured reliably; have construct validity and predictive validity; and be understandable and acceptable to experts and to consumers. (Ibid, p 272).

It is important that specifications such as these, regarded as basic requirements for the development of norm-referred tests, become essential characteristics of criterion-referenced assessment also. However, the difficulty still remains that:

> The establishment of standards and the defining of grade-related criteria in a criterion-based system is a hazardous undertaking .... What counts as an appropriate standard for a grade depends on the purpose, and purposes of academics, teachers and employers are not the same. Standard setting and the award of criterion-related grades are matters of judgement. (Power, 1986, p 275)

Despite the numerous problem areas to be confronted, Power still advocates further, gradual advancement towards C.R. assessment, but certainly never at the total exclusion of norm-referenced testing. Clearly, Power has endeavoured to produce an unbiased review, citing many of the advantages of C.R. assessment as well, yet his conclusions leave the reader with the realization that C.R. assessment cannot be implemented overnight (nor perhaps in a decade) or in all subjects, without a massive research input:

> If the impression has been created that what exactly is being called for by advocates of criterion-based assessment is unclear.... that it may not be possible to have a coherent, reliable and functional criterion-based system which is specific about what candidates have and have not achieved, that... was intended. If, as well, it seems that we may end up not far from where we started, that too is a real possibility. (Ibid, p 281).

The Ross Report (1986), has also recognised the problems and lack of research related to C.R. assessment and that many of its advocates do not fully understand these issues (p 58). What then, is the best available alternative to moderate internal assessment with, as an interim measure at least, and as a potential method for Group One subjects? Referring back to Table 2.1, two further options appeared as equally attractive to

teachers as suitable methods of assessment in Form 6. With about 20 percent of popular support each, option (b), Internal Assessment Moderated by Reference Tests, and option (c), Partial Internal Assessment in combination with an External Examination stand out as the only likely alternatives. The next most attractive option could gather only nine percent of support (namely, the current system).

Of these two options, the latter does not appear as a feasible system at Form 6 for three important reasons. Firstly, the recommendations of the Ross Report are for total internal assessment at Form 6. Since many of the over 400 Sixth Form Certificate courses have been developed with this in mind, it seems most unlikely that even a partial examination would be reintroduced. Secondly, if a partial examination was reintroduced in Form 6, the problems of teachers teaching to the examination syllabus would arise again, and the internally assessed component would tend to reflect this also as teachers would be loathe to utilize content that did not assist with the examination syllabus. Thirdly, it is likely to be viewed rather as a step backwards by the PPTA, who have fought long and hard to introduce internal assessment. Such a move may cause more conflict than progress, as in Queensland for example.

Perhaps the best alternative, given the current situation, is that of internal assessment moderated by reference tests. In some respects, this is similar to the status-quo, where School

Certificate Examination results have been used to moderate Form 6 courses. However, this method does not take into account the maturation of pupils' work habits from one year to the next, nor does it take account of pupils' selecting new subjects in their Form 6 year. In addition, it also demands the presence of an examination at Form 5, which, according to present policies under review, is to be removed once a suitable moderation scheme is in place at Form 6. However, despite these criticisms,the basic principle has worked well in maintaining comparability between schools, subject and year groups (Ross Report, p 64).

The advantages of professionally developed reference tests include a highly reliable and valid instrument, easy administration and marking, being relatively inexpensive for the school and leaving teachers free for teaching and preparation rather than having to act as moderators. The one major disadvantage would be the danger of teachers teaching to the reference test, in the same manner they taught to examinations. However, a suitable test format, in essence a cross between an aptitude and achievement test, and the use of group sampling techniques, have the potential to minimize this problem.

The final argument in favour of using reference tests is that there is already a reasonable research base to work from - the development and application of ASAT and earlier projects in Australia, achievement tests in Sweden and aptitude tests in the USA. Locally, there have been two major studies by Hulbert (1978) and Gilmore (1979), which made use of reference tests with promising results.

CRITERIA FOR AN IDEAL MODERATING TEST

Gilmore (1979), as part of her study, conducted an extensive search of the research literature as a survey of moderation tests in current use. She identified the following criteria as the most important features to consider when evaluating the suitability of a moderating test. Moderation tests should:

(1) be able to detect the arbitrary differences between the marking standards of different markers;

(2) be able to detect real differences in the calibre (level and distribution of ability) of groups of pupils (classes or schools);

(3) require relatively little time to administer;

(4) cover as wide a range of curricular and behavioural objectives as possible so that a broad spectrum of pupil ability may be assessed;

(5) not place undue 'examination stress' on the pupils;

(6) not be susceptible to cheating;

(7) have reasonable 'face validity', acceptability to teachers and pupils;

(8) reflect pupil skills which result from school study and are influenced by teaching quality;

(9) not be susceptible to being directly taught to, that is, it should not encourage 'coaching' by teachers [i.e. 'backwash' effects];

(10) not be based on prescribed syllabuses; and

(11) not encourage rote memorization of facts, formulae, standard procedures, etc. (Gilmore, 1979, pp 4-5)

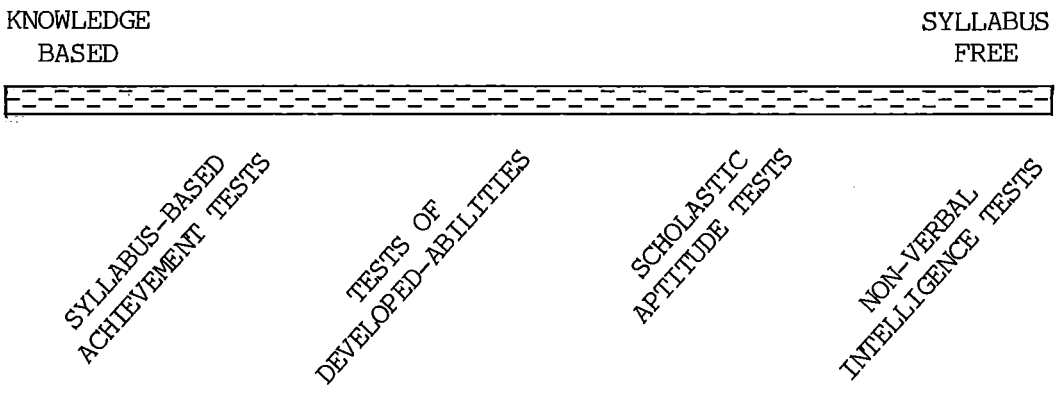
In accepting these criteria, they will be used to assess the relative merits of the reference tests under construction in this study.

TYPES OF TESTS SUITABLE FOR MODERATION

The nature and types of test materials available to act as a moderating instrument can range from the partial use of an external examination (e.g. School Certificate) to a commonly used verbal-intelligence test, such as TOSCA. Levine (1958) for example, has suggested that the different nature of tests can best be described using a "test-content" continuum. Levine distinguishes various tests by their "degree of subject-matter generality". That is, the extent to which different types of tests contain specific knowledge (i.e. achievement tests) or attempt to assess general school-related skills (i.e. aptitude tests). Along such a continuum, different types of potential moderating tests can be identified, as can be seen from Figure 1.

FIGURE 1

Classification of Potential Moderating Tests on a Continuum of Test-Content



Examples:

SC, UE, PAT-Mathematics (NZ) 'O'/'A' Levels (UK)	PAT Rding/Vocab, Hulbert, Gilmore (NZ) TEEP/ASAT (Aust) IOWA, SCAT, STEP (USA)	OTIS, TOSCA (NZ) APT, DAT, FACT, SAT (USA)	RAVEN'S STD PM, CULTURE-FREE/FAIR IQ TESTS

Clearly, the two extremes of the test-content continuum would be unsuitable as moderating instruments. The fate of external and school-based examinations has already been discussed at length, except to point out, that as examinations are largely knowledge-based, then by implication, they must be tied to very specific curricula. Therefore, as a form of assessment, the examination syllabus comes to dominate the curriculum to the extent that innovation at the classroom level is greatly inhibited. At the other extreme, non-verbal intelligence tests are virtually knowledge-free and, therefore, are not related to any particular curricula. The items are usually abstract in nature with little or no written language. Consequently, they do not reflect the kind of school-related tasks under consideration here. Although such tests still correlate reasonably highly with school achievement, their poor face and content validity must render them unsuitable for moderation purposes.

Moderation Through Tests of Scholastic Aptitude

Tests of scholastic aptitude (or verbal intelligence) are designed to measure a broad mixture of verbal, abstract and numerical skills (i.e. a pupil's aptitude for learning) which are considered fundamental in a variety of academic subjects (Elley and Livingstone, 1972, pp 100-1). Pupils who have difficulty developing these skills tend also to have trouble keeping pace with their peer group at school. This type of test has been used extensively in New Zealand schools (e.g. Otis and TOSCA[1]) as part

[1]Test of Scholastic Abilities

of the process of allocating pupils to courses and streams upon entry to secondary school (Matthews, 1983). This type of test has also proved popular overseas, especially in the USA, where it is used as a mass screening and selection device for entry to tertiary institutions (Keepes and Keepes, 1974).

Tests of scholastic aptitude have two important advantages as a moderating instrument when compared with achievement tests. Firstly, as can be seen from Figure 1, they are largely syllabus-free, and therefore, are not susceptible to "backwash" effects, thus allowing the classroom teacher greater flexibility to develop alternative curricula; and secondly, it has been shown that coaching has only small effects on a pupil's performance (e.g. Yates, 1953; Frankel, 1960).

The effectiveness of scholastic aptitude tests as a moderating instrument can be demonstrated from Gilmore's (1979) study. Included among her battery of tests was a specially constructed general aptitude test (GAT). As a predictor of class parameters the GAT produced a median correlation of 0.85 (range .94 to .79). In fact, the GAT predicted School Certificate English and Science as well as or better than the respective reference tests, and at only a slightly lower level for School Certificate Mathematics. When Gilmore conducted the multiple regressions it was found that the GAT formed a major component in nearly all the prediction equations.

A number of other studies have investigated the predictive validity of aptitude tests in relation to examination performance at a level equivalent to Form 5 (Schools Council, 1965 , 1966 , 1972; Skurnik and Hall, 1968; Nuttall, 1971; Elley and Livingstone, 1972; Hulbert, 1978). The two latter studies both reported median correlations for the prediction of individual scores of 0.55 (range .36-.81 and .49 - .76) between Otis and School Certificate marks, as well as median correlations of 0.44 and 0.52 (range .31 - .71 and .35 - .72) between the Differential Aptitude Test (Verbal Reasoning) and School Certificate marks. Using group discrimination procedures it should be expected that the correlations would improve significantly again. These results suggest that there is some potential for aptitude tests to act as a moderating device at this level.

However, if moderation is to be applied at the Form 6 level, as the Ross Report recommends, there is likely to be a related problem. The current retention rate at this level, although growing, is still only about 60 percent of the Form 3 intake due to the screening effect of the School Certificate Examination. Therefore, an aptitude test is unlikely to be able to discriminate significant differences between groups of pupils at a sufficiently sensitive level. In effect, the test is attempting to predict amongst a more homogeneous sample than that upon which it was developed. Several studies (e.g. Black, 1960; Elley and Livingstone, 1972 and Choppin, Orr, Kurle, Fara and James, 1973) have all identified this problem.

Another criticism of using scholastic aptitude tests for moderation purposes is that they fail to assess school-related skills, for example, perseverance, work habits and background knowledge. In addition, such tests are also insensitive to the quality of the teaching (Dunn, 1977, p 13) and lack much in face and content validity (Elley and Livingstone, 1972). Thus, perhaps in the same way that achievement tests were too closely tied to specific curricula, one could equally state that scholastic aptitude tests are too distant from the curricula, to the extent that teachers remain very suspicious of them. One question from Elley's (in progress) national survey on assessment policies asked teachers what type of test content they would most prefer if reference testing was to become policy. Only 1.5 percent of teachers favoured the use of a "scholastic aptitude or general intelligence" test.

Moderation Through Tests of Developed Abilities

What exactly are tests of "developed abilities?" Berkeley and Alford (1974), for example, felt that the easiest method of defining developed abilities is to describe what they are not. Thus, a test of developed abilities is not...

(1) a syllabus-based achievement test or a specific attainment test, such as the traditional public examination;

(2) based on any prescribed course of study (i.e. it is content- or syllabus-free); and

(3) a general ability or intelligence test (p 111).

Other researchers have defined the relative classification of the developed abilities construct in relation to other test

constructs. Gilmore (1979), for instance, viewed a test of developed abilities as a "compromise" between tests of achievement and scholastic aptitude, as presented earlier in Figure 1. As a potential moderating test, she regarded this type of test construct as a "balancing act" between that which can offer maximum curricula freedom while still relating to school-based activities. Similarly, while considering various options for moderation, Elley (1985, pp 13-14) identified an important and critical balance to aim at,

> ... if we want the flexibility to develop new locally relevant courses, and still aim for parity of standards - then we need moderating tests which fall half-way along the spectrum between traditional knowledge-based examinations, like School Certificate, and syllabus-free tests, like those which measure intelligence.

Thus, a suitable description of a test of developed abilities is that it attempts to measure general school-related abilities (e.g. general comprehension, interpretation and application of basic concepts and reasoning skills) within a discipline, but which do not include the recall of distinct facts, knowledge or procedures related to any particular syllabus or learning programme.

The major research thrust in relation to tests of developed abilities has taken place in Australia with the development of the Commonwealth Secondary Scholarship Examination (CSSE), the Tertiary Education Entrance Project (TEEP) and the Australian Scholastic Aptitude Test (ASAT). Unfortunately, not much can be gained from the research relating to the CSSE since its introduction in 1964. After reviewing the relevant studies, Thomson and Slee

(1975) concluded that due to a "lack of comparative data and the truncated nature of the samples studied", it was virtually impossible to make any definite claims concerning the test's predictive power.

The TEEP in 1969 and ASAT in 1970, were introduced in response to the growing dissatisfaction over external examinations and a demand for internal assessment procedures instead. ASAT was developped directly from the TEEP research and thus, should be considered as of the same origin. Only ASAT is referred to in the following studies.

The ASAT is a three-hour objective test of 100 questions, intended for administration at the Year 12 (or Form 7) level in Australian schools. Care is taken to avoid specific content of Years 11 and 12 syllabuses and courses of study. The questions are designed to measure, in the main, the abilities of comprehension, interpretation and reasoning within broad school subject categories. Some of the skill factors to be assessed include verbal ability, short term recall, interpretation of data, evaluation of data, reasoning ability, judgemental ability and visual skills. The Test Specifications (no date) stated that,

> ... its major use is that of scaling school assessment both between subjects and between schools. Consequently, its validity must be considered at all times in terms of its strength and robustness in carrying out this task. (p 4)
>
> The reliability of each edition of the test can be confidently presumed to be equal or greater than 0.90, measured by KR-20. (p 4)

Technical studies of past ASAT series by Bell (1973 and 1977) confirmed the test's soundness in these aspects.

A number of smaller regional studies have been reported (e.g. Sutherland, 1972; Otto, 1976; Rosenberg, 1976; McGaw, 1977; Queensland Board of Secondary School Studies, 1978 and 1978a) but four large, state-wide studies, often mentioned, provide a good cross section of the ASAT-related research (Lees, 1979).

The Public Examination Board of South Australia (1975) investigated the ASAT as a possible system of moderation. A total of five procedures for moderating internal assessment were evaluated: (1) ASAT arts subtotal; (2) ASAT science subtotal; (3) ASAT total score; (4) a raw external examination mark; and (5) a matriculation total (comprising an aggregate of rescaled examination marks). The validity criteria selected were the internal assessment marks of a pupil's achievement in each subject. The results revealed that the two non-ASAT procedures were the better predictors of the internal assessment marks. The best predictor was the raw external examination mark with a median correlation of 0.70 (range .48 to .84). In comparison, the ASAT total (the best of the three ASAT scores) produced a median correlation of only 0.37 (range .20 to .67). The South Australian authorities chose to ignore other important educational benefits (e.g. backwash effects and curricula freedom) and dismissed the ASAT proposal without further consideration.

In contrast, the Queensland Department of Education has reported evidence to support the use of ASAT as a moderation test. A study by McGaw, Warry and McBryde (1975) evaluated subtotal and

total scores from ASAT to predict school aggregate internal assessment marks. The total score, which rated consistently better than the subtotals, produced a correlation of 0.62, and 0.64 when replicated three years later. Additional evidence, from teacher feedback, showed that there has been a notable decrease in backwash effects and the teachers' assessments have improved with a more sensitive grading system over the three year period of the study.

A review of the moderation procedures in the Australian Capital Territory by Keeves, McBryde and Bennett (1977), focussed on issues relating to ASAT's validity using Higher School Certificate as a criterion measure. Among the questions asked was whether ASAT acted as a good reference test. The empirical results demonstrated ASAT to be a highly satisfactory method of moderating pupil achievement between secondary schools and between subjects (except for modern languages and art).

The final state-wide review came from Victoria, where MacKay and Fary (1979) analysed the possible roles of information from testing programmes in the transition from secondary to tertiary education. The data was based on a particularly large sample of 18 752 Year 12 pupils from 405 schools and 180 tertiary courses. The large study investigated several different moderating procedures including ASAT. The findings revealed that the ASAT: (1) had acceptably high reliability estimates; (2) mathematical and verbal sub-tests provided valid scores; (3) could differentiate significant differences between identifiable groups of pupils; and

(4) proved to be the best of several moderating tests under consideration because it displayed the greatest stability in predictive validity across subjects. Despite these encouraging results ASAT still does not play a part in Victoria's official assessment procedures at Year 12.

Three New Zealand based studies, including two major thesis investigations, have produced some promising results.

Evidence of the predictive validity from the widely used PAT Reading Comprehension test found that it predicted individual School Certificate English marks at 0.75 two years later and University Entrance English at 0.50 three years later (Elley and Livingstone, 1972, p 112). A more extensive New Zealand study was conducted by Hulbert (1978) in which he developed a test battery as a potential means of either scaling or moderating assessments for School Certificate. Included in the battery was an Experimental Moderating Test (EMT) comprising a series of comprehension passages based on general topics, along with an associated studies skills section requiring pupils to make appropriate notes to help answer questions about the passages. The EMT had approximately 35 multiple-choice items and was administered to a sample of 306 Form 5 pupils from ten Hamilton schools. The results, using the School Certificate Examinations as the criterion variable, produced a median correlation of 0.47 (range .41 to .66) with the EMT, based on an individual level of analysis. Hulbert conducted an exploratory factor analysis which showed that the EMT contained a very high verbal comprehension

component similar to the School Certificate English Examination.

In both these studies the analysis was based on the prediction of each individual's results, as against the prediction of class or subject group performances, which is the major task of a moderating test. Since the prediction of group performances is an easier task and usually produces higher correlations, then these results - especially that of the PAT Reading Comprehension test - look even more impressive. However, there has been one study which did focus on the prediction of group performances, specifically in relation to the issue of moderation.

This final piece of local research involved a Ph.D. thesis investigation at Otago by Gilmore (1979, also reported 1984). Tests of developed abilities were used as the basis for a series of moderation tests to be validated by predicting pupils' performances on the School Certificate Examination. Four parallel tests (16 in total) were constructed from item pools in each of three core subject areas - English, mathematics and science - as well as an additional test of general scholastic abilities. The test items avoided specific syllabi in the Form 5 curriculum, but did assume general knowledge and recall of basic facts in the respective courses at the Forms 3 and 4 level. Gilmore described the construct of developed abilities as a "compromise" between tests of achievement and scholastic aptitude, a test that,

> ... would employ 'syllabus-free' material which is substantially independent of any prescribed course of study, but would assess skills or abilities developed after the pupil has undertaken a particular course of study. The acquisition of these 'developed abilities' would ... be retained longer by the student than specific

> items of factual knowledge. These skills ... would have greater flexibility in developing curricular content ... more acceptable face validity ... and would reflect both teacher and student effort.
> (Ibid, p 69)

Her research hypothesis was to assess the performance of the respective moderating tests (singly and in combination) as optimal predictors of class parameters (M and S.D.) for a number of School Certificate subjects.[1] The tests were administered to 922 fifth formers from five Otago secondary schools, early in the third term. The main findings are listed below:

(1) generally, the four moderating tests provided very good group predictions of the School Certificate class means. The median r's were 0.73 for the English test (range .71 to .94), 0.86 for the mathematics test (range .44 to .93), 0.87 for the science test (range .59 to .89) and 0.85 for the general aptitude test (range .71 to .94). Two of the three subject tests predicted their respective School Certificate subjects at a higher level than any other test. The English test predicted School Certificate English with an r of 0.94, the maths test correlated 0.93 with School Certificate Mathematics and the science test 0.87 with School Certificate Science. Two other tests, general aptitude and science, predicted S.C. Science slightly better at 0.91;

(2) in half the subjects (English, Science, Geography, History and Typewriting) the general aptitude test produced the highest or equal highest group r;

---

[1] The School Certificate subjects were English, Mathematics, Science, Geography, History, Biology, Economic Studies, Technical Drawing, Typewriting, A.L. French.

(3) for those subjects in which multiple regression equations were formulated (English, Mathematics, Science, Geography, History and Economic Studies) the general aptitude test proved to be a very significant predictor of the criterion; and

(4) a combination of selected moderation tests (multiple-R's) predicted School Certificate class parameters (M and S.D.) at a slightly higher level than the simple correlations. The median R for the class mean was 0.92 (range .71 to .96), and for the class standard deviation 0.72 (range .59 to .83).

Gilmore's major conclusion was that while a general underlying factor as measured by the scholastic aptitude test was very evident, it was also clear from the regression analyses that specific abilities could partially explain pupil performance in various School Certificate subjects. These abilities were largely developed through "school-experiences" and expressed particularly in the predictor disciplines of English, mathematics and science (Gilmore, 1984, pp 72-3).

In view of the research and related findings under review here, tests of developed abilities must be considered as an eminently suitable instrument for moderation purposes. This approach not only acocunts for the main evaluation criteria of discrimination power, face and content validity and minimization of backwash effects, but is also inexpensive and does not produce a heavy administrative burden on individual schools (assuming, of course, that an outside agency will be responsible for all facets of the testing).

SUBJECT AREAS IN MODERATION TESTS

The range of subject areas which have formed components of moderation tests is fairly restricted with several common areas being shared by different tests. For example, the three tertiary entrance tests developed in Australia shared common subject disciplines. They included the sciences and humanities in the CSSE; the sciences, humanities and social sciences in the TEEP; and the sciences, humanities, social sciences and mathematics in the ASAT.

A local study by Elley and Livingstone (undated), collected evidence indicating that the average School Certificate Examination mark in English plus Mathematics or Science (the EMS composite score) was a very accurate predictor of pupils' performances in a range of other subjects. A 1966 study using a sample of 3537 candidates demonstrated that the EMS score correlated very highly with School Certificate four-subject total (r=0.94), certainly higher than School Certificate English, Mathematics and Science separately (all r = 0.87). The study also showed that the EMS score correlated very highly with School Certificate English, Mathematics and Science separately (all r = 0.87). The study also showed that the EMS score correlated very well with a wide range of School Certificate subjects (except Practical Art). The correlations ranged from 0.43 to 0.87, with a median of 0.64, the smaller correlations tending to associate with those subjects that have a significant practical component (e.g. Typewriting and Woodwork). In general, however, the results indicated that the EMS composite score would be sufficiently sensitive to detect real differences in the level and range of abilities across different groups of pupils.

Moderating tests used overseas also tend to focus on the core subjects. In Sweden, for example, achievement tests were developed in Swedish, maths, science and two foreign languages. The New South Wales education authorities are now moderating in only two subjects, English and mathematics, after earlier policies had demanded moderation in all 30 School Certificate subjects at year 10 (Elley, 1985, p 13). Of the two major studies conducted in New Zealand, Hulbert (1978) used comprehension passages from the sciences and social sciences, while Gilmore (1979) developed tests in English, mathematics and science, as well as general scholastic abilities.

In view of the excellent predictive capabilities demonstrated by Elley and Livingstone (Undated) and Gilmore (1979) using English, mathematics and science to predict a wide range of School Certificate Examination marks, it was decided to continue and extend the idea of using core school subject areas as predictor variables. In addition to the three subject areas mentioned, a social studies test was developed to be assessed in its own right as a predictor and to help redress the mathematics/science emphasis in the test battery.

VOCABULARY AND COMPREHENSION

Vocabulary refers to the whole range of words known or used in a particular school subject area, such as English or science. None of the words used in the vocabulary tests were specific to the respective syllabi at the Form 5 level. Each

word was presented in "context" form as a complete sentence. Some examples from the tests are illustrated below...

ENGLISH EXAMPLE: The writer's epilogue was particularly impressive.

SCIENCE EXAMPLE: The doctor stressed the need for a balanced diet.

SOCIAL STUDIES EXAMPLE: One of our basic commodities.

It might be claimed that the use of vocabulary is really "word knowledge", and that knowledge by definition is not considered part of the developed abilities construct. However, by not focussing on the current Form 5 curriculum, only previous basic knowledge from earlier courses in being assumed, and associated with acquiring this basic knowledge are all those skills and school-based abilities that form the basis of the developed abilities construct. In addition, it may be argued that vocabulary could be susceptible to "coaching" by more academically-oriented schools; however this seems unlikely since the potential number of words that could be chosen for testing within each subject area would be so large that such practices would prove to be of little or no advantage.

Comprehension refers to the power or ability to understand written prose as it relates to a particular school subject area, such as social studies. The passages were of a general nature and were not tied to the specific content of the respective syllabi at the Form 5 level. Because of their length examples of the passages will not be illustrated here, but can be seen in full in Appendices D and E.

The use of vocabulary and comprehension to measure verbal and school-related abilities has a fairly extensive history. Vocabulary, in particular, was a powerful discriminating component in the early stages of the intelligence testing movement and is enjoying a modern-day resurgence as researchers have realized its importance, both singly and in relation to comprehension. Comprehension, of course, has always been considered an important and logical means of assessing not only school-related success, but also, as a basic requirement for living in an increasingly literate world.

Vocabulary knowledge has been a major component of many intelligence tests and the relationship between the two has been shown to be a strong and consistent one (e.g. Terman, 1918; Elwood, 1939; Lewinski, 1948; Dupuy, 1974). Anderson and Freebody (1981), having reviewed 11 relevant studies, reported correlations ranging from 0.71 to 0.98 between vocabulary sub-tests and total tests from various intelligence and achievement tests. A similar relationship is evident between the New Zealand PAT Vocabulary test and TOSCA, with correlations ranging from 0.74 to 0.86 over six grade levels (Reid, Jackson, Gilmore and Croft, 1981).

Another strong relationship is that between vocabulary knowledge and comprehension. Davis (1944 and 1968), for example, identified a major vocabulary factor ($r = 0.80$) after factor analyzing nine comprehension tests. Again, Anderson and Freebody (1981), reported a range of factor loadings, from four

further studies, that correlated from 0.41 to 0.93 with vocabulary tests. From the PAT Vocabulary and Comprehension tests, correlations were reported as ranging from 0.79 to 0.86 (Elley and Reid, 1969), once again, consistent with the overseas findings.

On the basis of these findings, attempts have been made to determine exactly what vocabulary measures. One theoretical position - that vocabulary measures verbal aptitude - argues that "people of high verbal ability are literally faster than other people at elemental verbal encoding and recoding operations" (Anderson and Freebody, 1981, p 82). The authors believe that:

> A person who scores high on [a vocabulary] test has a quick mind. With the same amount of exposure to the culture, this individual has learned more word meanings. He or she also comprehends discourse more readily than the person who scores low on a vocabulary test. [Thus] ... persons with large vocabularies are better at discourse comprehension because they possess superior mental agility. [Therefore] ... vocabulary test performance is merely another reflection of verbal ability (Ibid, p 81).

In New Zealand, vocabulary knowledge forms part of the Progressive Achievement Test series which is used very widely in schools from Standard 2 up to Form 4. However, the use of vocabulary and comprehension in reference tests for moderation purposes has not been attempted before. Similarly, there has been no research on their suitability for measuring the construct of developed abilities within specific subject disciplines.

FORMAT OF TEST ITEMS

The secondary aim of the study sets out to compare the relative merits of alternative types of items. The three item formats selected for investigation include the multiple-choice, open-ended or "supply" (sometimes referred to as the short-answer) question and the cloze procedure.

The Multiple Choice Item

This is probably the most versatile and rigorously objective item available to the tester. But it is also the most difficult to construct. Multiple-choice items consist of a stem (i.e. a question or statement) and a series of options or alternatives, one of which must be the key or answer, the other options acting as distracters. It is the most widely used of the objective items (Hudson, 1973, p 125) and is particularly popular in commercial tests (e.g. the PAT series) because of its precision, total objectivity during marking and suitability for item analysis.

A major area of debate concerning multiple-choice items is the number of options that should be provided. Typically, four or five options are offered (certainly this is the case with commercially produced tests), but the use of only three options has proved popular with classroom teachers and researchers. Numerous studies (e.g. Tversky, 1964; Costin, 1970; Grier, 1975; Swanson, 1976; Straton and Gatts, 1980) have claimed that long tests of three option multiple-choice items are superior to short tests of four option items.

However, a reanalysis of the two latter studies by Duncan (1983), has resulted in several major criticisms of the methodology employed and of their interpretation of the results. After his reanalysis of the results it was concluded that:

> In addition to the reported superior discrimination values, the superior average percentage correct data indicate that tests with four-alternative items are more psychometrically sound than tests with three-alternative items. (Ibid, p 291.)

Another study by Ramos and Stern (1973), investigated the differences between four and five option items in specially constructed language tests. From the results the authors concluded that:

> The use of four-rather than five-choice items ... should result in gains in test development of efficiency and lower cost per item... however... these gains in efficiency must be traded off against [small but significant] losses in test reliability and item discrimination (Ibid, p 309).

Perhaps the major consideration as to whether four or five option items should be selected is the length of the tests. Since two important features of a moderation test are to keep testing time to a minimum while maximizing group discrimination, it would be sound reasoning to use short tests with five-option multiple choice items.

The Open-Ended (or Short Answer) Item

The open-ended item probably measures something slightly different from objective type items which supply possible answers or further context, these often acting as recognition cues for the examinees. Thus, for open-ended items the pupil must not only

recall the appropriate material, but also write the answer down in their own words (Brown, 1966; MacIntosh and Morrison, 1969).

Sax (1980, pp 125-6) lists the advantages of the open-ended item as: (1) the relative ease of construction; (2) the problem of guessing is eliminated; and (3) the range of items sampled is improved since less reading time is required. The disadvantages include the difficulty of marking and the tendency to measure mainly rote objectives. The major difficulty in the marking procedure lies with the simple fact that, as each answer is written down there is an opportunity for confusion to occur, either through a poorly written question which does not specify the exact form or degree of answer required, or through "border-line" answers the examiner has not anticipated yet which seem quite logical (Willmot and Hall, 1975, p 39).

Thus, the main point of interest in relation to the open-ended item format (notwithstanding the lower reliability due to the subjective nature of the marking) is whether it will be any less efficient than the multiple choice item. The research in this area (e.g. Heim and Watts, 1967; Traub and Fisher, 1977; Ward, 1982; Frary, 1985) was in some parts contradictory, however, one general indication was that while the performance of the open-ended item may be suitable depending on the requirements of different subject areas (e.g. English versus mathematics), other important considerations - such as the extra marking time needed in comparison to the machine scoring of multiple-choice items - might weigh against its use in a national moderation test.

The Cloze Procedure

For some unknown reason the cloze has been consistently ommitted from texts of educational measurement (Baldauf, Jr, 1980). However, it has long been considered a useful "tool", and also a focus of interest in its own right with educational researchers. In addition, its popularity with classroom teachers has accelerated markedly over the last decade or so.

The cloze procedure involves the systematic deletion of words from a prose passage. Pupils are then required to replace the missing words using the context of the passage (and sometimes multiple-choice options). Briefly, the construction technique involves deleting every nth word (for young children it may be every 10th word, but for older pupils the pattern is usually every 5th; some researchers have investigated random deletion as well, for example, Helfeldt, Henk and Fotos, 1986) after leaving the first line or sentence of the passage intact. Standard length blanks are inserted for each deleted word for obvious reasons.

The cloze question has received much attention from researchers in the areas of measurement of reading comprehension (e.g. Jenkinson, 1957; Bormuth, 1969; Anderson, 1974; Marandos, 1974; Reid and Hughes, 1974; Cunningham and Tierney, 1979; Dupuis, 1980; Hosseini and Ferrell, 1982), as a teaching device (e.g. Rankin and Dale, 1969; Bortnick and Lopardo, 1973) and in assessing readability (e.g. Taylor, 1953; Slane, 1968; Bormuth, 1967; Elley, 1967 and 1969; Entin and Klare, 1978).

Despite this wealth of research there are still many areas of debate. One of these is over what exactly the cloze attempts to measure (e.g. Alderson, 1978). Elley (1976a) has suggested an explanation about the likely mental processes that are involved. There are three main factors:

(1) knowledge of the subject of the passage;

(2) understanding of language structures; and

(3) ability to draw inferences.

> We read with most understanding if we know the subject well, are familiar with the conventions of language used, and apply some reasoning processes in the act of reading (Ibid, p 56).

Elley's analysis, influenced in part by Smith's (1971) psycholinguistic theory of reading, provides some useful theoretical considerations in relation to the construct of developed abilities. "Knowledge of subject", "understanding of language" and "ability to draw inferences" are all highly important, school-related skills that form part of the developed abilities construct, this being the basis for the development of the reference tests.

Another area of debate concerns the scoring procedure of the cloze. In the past, the general recommendation has been that only exact replacements (i.e. the original word omitted from the passage) be marked as correct - ignoring spelling errors. Thus, in many related studies synonym replacement has not usually been considered since the research evidence has indicated that it does not alter significantly the rank order of pupils' scores, while tending to reduce the reliability of the test and add greatly to

the marking time (e.g. Bormuth, 1965 and 1968; Hargis, 1972, McKenna, 1976; Elley, 1976a and 1977). A recent study by Henk (1981), in part, assessed the relationship between exact and synonym replacement. Spearman rank-order analysis produced a median correlation of 0.75 (range .71 to .75) between the two scoring procedures. Although this result was somewhat lower than typically observed (.85 to .95), it still provides another indication as to the similarity in outcome between the two scoring methods.

The use of the cloze procedure as a measure of developed abilities and/or as an alternative item format in moderation tests appears to have been the subject of no research. However, the qualities of the cloze are in many ways just as impressive as those of the multiple-choice item as a convenient and completely objective measurement procedure (Bormuth, 1969). In addition, Bormuth (1967 and 1968), Rankin and Culhane (1969), Entin and Klare (1978), Elley (1984) and Ratnamalar (1986) have demonstrated that cloze test percentage scores are highly equivalent to multiple choice reading comprehension scores. On this basis, it seems highly likely that the cloze procedure possesses very good potential as an alternative assessment technique for moderation purposes.

Why Compare Alternative Item Formats?

The multiple-choice, and, to a lesser extent, the cloze, are both examples of objective item formats[1], in that there is only one predetermined answer for each question, whether it is

[1]Strictly speaking, the cloze is only semi-objective, since there can be more than one correct answer, depending on the marking scheme employed.

answering "true" or "false" or choosing from several options. The open-ended item, on the other hand, demands that the pupil supply an answer, without additional cues, and with no opportunity to succeed through blind guessing (Nitko, 1983, p 159).

Many of the advantages associated with objective test formats are also important considerations in the development of a moderation test. Ease of administration and fast accurate marking, for example, are two essential features if the results from reference tests taken early in the third term are to be centrally marked and analyzed, and then returned to the schools in time for the allocation of grades and final course assessments. Other advantages cited in various measurement textbooks (e.g. Thorndike, 1971; Sax, 1980; Nitko, 1983) include the following:

(1) specific and wide content coverage;

(2) precise problem posed;

(3) greater homogeneity in specifying test content;

(4) pretesting of items and determination of difficulty levels;

(5) produces high test reliability

What about the disadvantages associated with objective item types? MacIntosh and Morrison (1969, p 13) listed three common charges made against them, namely that such items:

(1) cannot test written expression or a candidate's ability to develop an argument;

(2) can all too easily test only factual recall or simple understanding of facts; and

(3) may encourage candidates to guess the answers to questions.

Certainly, the first criticism is a valid one, to the extent, that it would be rather difficult to convince many English teachers of the face and content validity of an English test without assessing written expression. Obviously, written communication is important not only to English teachers, but also to the individiual's future success in life. The Australian Council of Educational Research has addressed the same criticism in relation to ASAT - despite the well documented research on the relatively poor reliability of essay marking - by including an essay component in the most recent series of ASAT. Gilmore (1979) also recognised this requirement and included an essay test amongst her test battery. Although the essay failed to contribute to the multiple-R prediction of School Certificate English, it did produce a simple correlation of 0.52 with the same subject. It would seem likely that this aspect of assessment deserves further investigation.

The second criticism is that the use of objective items will lead to an over emphasis on the recall of facts and rote memorization. While this may seem the case on the surface, a closer examination of the techniques for writing objective items suggests otherwise. In fact, Sax (1980, p 101) listed as one of the advantages of objective items their "... great versatility in measuring objectives from the rote knowledge level to the most complex level." Brown (1966, p 5) stated in a similar fashion that such items measure "factual knowledge of the theme, comprehension of the principles involved and/or ability to deduce from these

principles". The Educational Testing Service (1972) has also provided a sound argument as to the capacity of the objective item for assessing intellectual skills at a higher level than basic rote memorization of facts. Sax (1980) has summarized a number of techniques available for writing items in formats that measure the higher cognitive skills as described by Bloom (1965) including the presentation of items in a unique situation; use of analogies to measure relationships; use of novel pictorial materials to measure principles; identification of assumptions and analysis of criteria; identification of relationships among similar topics; selection of appropriate examples of principles or concepts; and the interpretation of data in tables and diagrams (pp 108 - 113).

The final criticism relates to the lesser problem of objective items encouraging candidates to guess the answers to questions (McGawe, 1974 and Rowley, 1974). A multiple-choice item, if constructed with a set of good plausible distractors using four or five options, is unlikely to be susceptible to guessing beyond the chance factor (i.e. 1 in 4 or 1 in 5). A problem can arise when items are constructed with two or three very unlikely distractors, thus turning the item into a 50:50 choice between the only two plausible options remaining, but there are several procedures available to counter such problems (e.g. Gulliksen, 1950; Aiken Jr, 1965). In addition, since the data was analyzed on a group rather than individual basis, using a reasonable sized sample of pupils and items, the effect will be for such chance factors to balance out within any single class group.

Why, then, is it necessary to compare different item formats? Traditionally, the multiple-choice item has been the dominant question format, especially with professionally produced tests. It has been somewhat less popular amongst classroom teachers because of the technical demands and time necessary to construct satisfactory items. But in relation to a national moderating test, teachers may fear that the use of multiple-choice questions to the almost total exclusion of other types of questions would give some pupils an unfair advantage. In addition, different subject teachers may have a preference for one type of item format over another, for example, the objective nature of mathematics may lend itself more closely to the multiple-choice item than an open-ended item (e.g. Forsyth and Sprott, 1980). On the other hand, the requirements of English which is far less objective in nature, would differ from those of mathematics or even science.

As part of a national survey on issues in assessment policy (Elley, in progress), one question asked teachers for their preference regarding item format if reference tests were, in fact, to become policy in the short or long term. Of the 460 teachers who responded to this question an overwhelming majority of 86.1 precent favoured a "variety of question types", while only 4.2 percent and 3.3 percent preferred the exclusive use of multiple-choice or open-ended/essay type questions respectively. This trend applied across all the main subject areas of English, mathematics, science and social studies related subjects. This trend had also been noted in an earlier study of moderation

by Hulbert (1978, p 138):

> The multiple-choice item as used in the construction of the EMT is technically superior to other forms of test item yet has not won wide acceptance among New Zealand secondary teachers. It is perhaps desirable that parallel reference tests using different types of item be used in further studies on moderation so that the relative effectiveness of the test forms may be noted.

Obviously, there has been and still is an urgent need to investigate this aspect of test construction. As well, the almost exclusive use of the multiple-choice item should be re-considered. Hopefully, the current study may be able to contribute by focussing on the essay, open-ended and cloze procedure to assess their efficiency and technical aspects as a moderating test. There has been no previous research in this area.

STATEMENT OF THE RESEARCH AIMS

On the basis of the findings from the literature review, two specific research aims have been formulated:

(1) to develop and validate a series of moderating tests, based on the underlying construct of developed abilities, in four core subject areas (i.e. English, mathematics, science and social studies). The tests are expected to predict, at a sufficiently sensitive level for moderation purposes, class parameters (i.e. mean and standard deviation) in the corresponding School Certificate Examinations; and

(2) to compare the relative merits of two alternative item formats, the open-ended and cloze procedure, with the technically proven multiple-choice question, as is traditionally used in reference tests for moderation and subject purposes.

The tests were administered at the Form 5 level incorporating a totally random multi-matrix sampling procedure. In conjunction with the first aim, it was also decided (using the various developed abilities tests and an additional test of general scholastic aptitude) to conduct multiple regression analyses to provide optimal predictions of the respective class parameters. For the second aim, similar analyses were conducted to assess the relative contributions of the respective item formats as a moderating test. Finally, the essay was evaluated in analyses relating to the School Certificate English criterion only.

CHAPTER III

METHOD AND CONSTRUCTION OF TESTS

SAMPLE

Four state, co-educational secondary schools from the Christchurch district were invited by letter (see Appendix A) to participate in the study. They represented a range of SES levels, and one of the four was located in a small semi-rural town. The only restriction placed on the selection of classes within the schools was that the pupils in all classes participating should be doing at least four School Certificate subjects. Each school was asked to offer at least four classes that would provide a wide range of abilities. The details of the final sample were as illustrated in Table 3.1.

TABLE 3.1

A Breakdown of the Sample by School, Total Number of Pupils and Number of Classes

School	N of Form 5 Pupils	N of Classes
1	121	5
2	113	5
3	115	5
4	101	4
Total: 4	450	19

DEVELOPMENT OF THE TESTS

This section will focus on the detailed aspects associated with the rationale underlying the tests, the general preparation involved and a description of the final test format.

Three of the four subject areas - English, science and social studies - were each assessed by separate vocabulary and comprehension tests. For administration purposes, the three separate vocabulary tests were combined together into a single Vocabulary Test booklet, as were the comprehension tests into a single Comprehension Test booklet (see Appendices B to E). The Mathematics Test, General Scholastic Aptitude Test and the Essay Test were all administered separately because either the subject area or the aspects attempting to be assessed were unsuitable as measures of vocabulary and/or comprehension.

Two parallel test forms were developed for each of the four subject areas. Firstly, parallel forms were constructed for use with multiple-matrix sampling (see p 72) primarily as a means of reducing testing time, while still ensuring an adequate coverage of general subject knowledge and behavioural skills. Secondly, parallel forms were also constructed to enable a direct comparison between the different item formats. Therefore, the question stems, stimulus material and prose passages were standard across parallel test forms for the respective subject areas, with only the item format changing (i.e. multiple-choice versus onen-ended or cloze, see pp 79-80).

Rationale Underlying the Moderation Tests

For the purposes of this study developed abilities refers to a common set of skills, aptitudes and background knowledge which are generated by those school-related experiences underlying all academic disciplines, no matter how diverse the actual curricula content. The acquisition and subsequent expansion of these developed abilities over a long time means they are generally more permanent, and therefore more fundamental to academic learning than memorization of specific facts and other curriculum knowledge. It is also assumed that a good teacher's main objective is the development of these skills which are partly independent of the content taught, and therefore less syllabus bound.

Gilmore (1979, p 70) identified what she considered to be the most important of these developed abilities, that is, the basis upon which tests being used for moderation purposes - including those in the current study - should be constructed:

(1) an ability to comprehend, interpret and apply basic concepts in each of the subject areas;

(2) an ability to reason logically in that area;

(3) an ability to handle data presented in quantitative, graphical and symbolic form; and

(4) an ability to communicate correctly and fluently in writing.

Other criteria related to the development of the tests were that any material required to answer questions be provided in the tests (avoiding specific recall) and it was assumed that the pupils would possess an "elementary knowledge" of basic facts in each subject area which had been acquired before and during their

third and fourth form years. The skills measured by the tests could be similarly related to Bloom's (1965) definition of "higher level skills".

In accepting these criteria, the assumption has been made that "developed abilities" is the most suitable test construct for the purposes of moderation. The basis for this support came from some promising research conducted locally (Elley and Livingstone, 1972; Hulbert, 1978 and Gilmore 1979) and in Australia (i.e. the development of ASAT). Syllabus-based achievement tests were not considered because of their restrictive effect on the curriculum, while the use of scholastic aptitude tests - although more promising as a potential moderating device - failed on the criteria of face and content validity as would be perceived by subject teachers. However, it was decided to include a general scholastic aptitude test on the basis of Gilmore's (1979) results, firstly as a comparison for the developed abilities tests, secondly, as a likely important component in the generation of a series of multiple regression equations, and finally, as an area of interest in its own right.

The general focus on using the core school-subjects of English, mathematics and science to predict other School Certificate subjects for moderation purposes (Elley and Livingstone, undated and Gilmore, 1979) has indicated this to be a highly suitable scheme. In adopting this approach, it was also decided to investigate and extend the scheme by including

a social studies test - with greater face and content validity - to predict such subjects as S.C,Geography, History and Economics.

Preparation of the Moderation Tests

The first stage of preparation was to assemble sufficient test materials from which the test items were selected using the criteria of developed abilities as defined earlier (p 66). Some items were taken from Gilmore (1979), others from a number of standardized achievement and aptitude tests from the USA, a few from local unpublished tests, while several items were constructed specially for the study by the researcher.[1] Items which proved more difficult for the researcher in deciding on their inclusion were also considered by a post-graduate colleague and/or the thesis supervisor for a consensus decision. Thus, it was on this basis that item pools were developed in the four subject areas of English, mathematics, science and social studies (comprising geography, history and economics), as well as scholastic aptitude, in readiness for the construction of the trialling tests.

Pilot testing was conducted late in the second term (August 1986) in a Christchurch State co-educational high school, not included in the sample proper. Because of a limitation in time and resources only about half the items from each of the Vocabulary and Comprehension pools were trialled. None of the items in the Mathematics or GSAT pools were tested since these were mainly

---

[1] A full list of the source materials from which items were selected appears in Appendix J.

revisions of earlier efforts by Gilmore (1979) which had already passed through a rigorous testing programme. In addition, only the multiple-choice format was trialled because of the need to check the effectiveness of the distractors used in this question format. Appropriate stimulus material for the open-ended and cloze formats were just as easily ascertained from the results of the multiple-choice format.

The criteria used to select items for pre-testing were firstly, any original items constructed which had undergone no trial testing at all, and secondly items which had been tested previously, but whose difficulty level was uncertain. Then, finally, equal numbers of English, science and social studies items were arranged to produce parallel forms of a Vocabulary and Comprehension Test (four in total). The multiple matrix sampling procedure (see next section) was trialled as well, on a total sample of 94 Form 5 pupils (the numbers completing each of the four forms ranged from 46 to 48). The total testing time was 45 minutes.

Each question was analysed using a simple item analysis procedure. Information was generated on the difficulty level, the discrimination index and the performance of each item's distractors. The statistical analysis revealed that there was an approximately equal percentage of Vocabulary and Comprehension items spread over a range of difficulty and discrimination levels, as presented below in Table 3.2.

TABLE 3.2

A Breakdown of the Range of Difficulty and Discrimination Indices for the Vocabulary and Comprehension Pilot-Tests.

Index Level	Difficulty		Discrimination	
	Vocab.	Comp.	Vocab.	Comp.
Greater than 0.70	10.0%	8.5%	10.0%	5.0%
0.55 - 0.69	26.7%	27.7%	22.0%	32.5%
0.40 - 0.54	38.3%	29.8%	30.0%	32.5%
Less than 0.40	25.0%	34.0%	38.0%	30.0%
Range	.23-.85	.20-.81	.23-.77	.23-.77

For the Mathematics tests Gilmore (1979) reported that items were drawn from each difficulty index level (reading top to bottom as set in Table 3.2) in the following ratio - 10:10:9:5. There was no data provided about the discrimination indices.

The items were evaluated on these statistical factors as well as on logical aspects such as content validity. The overall results indicated that some words and passages were too difficult and these were dropped from the testing programme. Other items were excellent, while most required some minor revision. It was expected that nearly all of the revised items would perform at an improved level. The pilot-testing programme proved to be a most worthwhile exercise in the shaping of the final test items.

Preparation of the final tests started with the allocation of items by content specification to one of two parallel forms for the Vocabulary and Comprehension tests (English, science and social studies) and the mathematics tests to ensure comparable difficulty. A reanalysis of the item-analysis data for the Vocabulary and Comprehension items and Gilmore's (1979) analysis of the mathematics items, indicated that a suitable proportion of items were being selected at appropriate difficulty levels (Shoemaker, 1973). At this stage, a check was made of the 1986 Form 5 syllabus for each of the four subject areas under consideration, to ensure that none of the items encroached specifically onto the course content at this level.

Identical test content was then transformed into open-ended and/or cloze versions. For the Vocabulary and Mathematics tests the open-ended item format was used exclusively. Generally, there was no problem in the transformation from the multiple-choice format, although some items required re-writing and a couple were omitted because they did not fit easily into the open-ended format. For the Comprehension Test a mixture of cloze and open-ended formats were employed. The open-ended format was used for poetry, analysis of graph data, and map work because the stimulus material was un-suitable for use with the cloze procedure. The remaining material-all prose passages - was assessed using the cloze format. Finally, single forms of the GSAT (in multiple-choice format) and Essay Tests were constructed, and separate answer sheets were produced for use with all multiple-choice tests.

Multiple-Matrix Sampling (MMS)

The concept of MMS, in broad terms, refers to the situation where the total sample is not given the same test (or set of items) to sit. In the traditional sampling technique the total sample is administered all the test items. Since moderation only requires information relating to group parameters the individual data associated with the traditional sampling technique is not relevant. Several studies have shown the estimated mean and variance for a random sub-sample of pupils as being equivalent to those obtained from a total-sample administration (Lord, 1962; Plumlee, 1964; Shoemaker, 1970a and 1970b).

A number of important advantages are associated with the use of MMS, especially in relation to the criteria employed to evaluate the use of a suitable moderating test, namely that:

(1) the amount of testing time per pupil is significantly reduced by the use of two, three or four parallel tests;

(2) the use of parallel tests also allows for a wider coverage of subject content and behavioural objectives by which a greater range of school-related abilities are assessed;

(3) "examination stress" is reduced since pupils are answering different test items and individual comparison is not directly made. Rather the emphasis is on "group" performance;

(4) the random distribution of parallel tests minimuzes the opportunity for cheating to occur during testing; and

(5) administration of the tests is not a major burden for the individual schools.

For these reasons, it was decided to include the MMS technique in the current study as the most suitable sampling method available for moderation purposes.

Description of the Reference Tests

Vocabulary (English, Science and Social Studies).

These tests attempted to measure pupils' knowledge of general words selected from the school-subject areas of English, science and social studies. The words relating to each subject area were set out in three distinct sections (see Appendices B and C). There were a total of 36 questions, 12 per subject area, which had to be completed within 15 minutes. Each word was underlined and presented in sentence context form, thus:

ENGLISH EXAMPLE: Consider the derivation of this word.
A meaning
B origin
C spelling
D ambiguity
E pronounciation

SCIENCE EXAMPLE: Endothermic reactions...
A contain heat
B absorb heat
C reflect heat
D transmit heat
E involve no heat

SOCIAL STUDIES EXAMPLE: Does this reveal your doctrine?
A discipline
B diversity
C strategy
D medical opinions
E belief system

This measure of vocabulary knowledge was designed to provide a strong indication of linguistic ability, verbal aptitude and reflect a broad subject knowledge in the respective subject areas.

Comprehension (English, Science and Social Studies).

These tests attempted to measure pupils' understanding of a series of general prose passages selected from the school-subject areas of English, science and social studies. There were a total

of nine passages grouped according to the respective subject areas, and a 50 minute time limit to complete the test. One of the science passages involved the interpretation of a graph, while one of the social studies passages utilized map reading skills. The English passages included a poem and a newspaper editorial to be analysed. Copies of the Comprehension Test can be seen in Appendices D and E.

These measures of comprehension were designed to provide a strong indication of the pupils' understanding, and therefore, their ability to analyse and interpret from different prose forms in the respective subject areas. The assumption here is that pupils with greater familiarity of a subject will be able to analyse and interpret knowledge better than those who are less familiar with it. This appears as a perfectly reasonable assertion to make.

Mathematics. These tests attempted to measure skills relating to general mathematical concepts. The tests were developed in part from the work of Gilmore (1979), and was based on the following criteria:

> (1) elementary arithmetic, algebraic and geometric computation and concepts;
>
> (2) interpretation of graphs and tables;
>
> (3) ability to generalize or make rules which satisfy many conditions; and
>
> (4) ability to reason mathematically, and solve problems of a general mathematical nature, using money and in a novel situation in which "nonsense" data was supplied. (Ibid, pp 73-4)

In contrast, the sort of criteria not included in the tests were those skills or degree of skill as specified in the Form 5

Mathematics Syllabus. Instead, the range of skills being assessed reflects the manipulation of elementary knowledge built up over earlier courses, rather than the advanced knowledge encountered at the Form 5 level.

There were 25 questions to be answered within a 25 minute time limit. Copies of the Mathematics Test can be seen in Appendices F and G. Some example questions from the tests are presented below:

(a) Which of the following represents the number which is 7 greater than 11?
A $7 > 11$
B $11 > 7$
C 11 x 7
D 11 + 7
E 11 - 7

(b) A family saves 5% of its monthly income. If the monthly income is increased from $600 to $650, by how much are the monthly savings increased?
A $ 5.00
B $ 2.50
C $ 1.75
D $ 1.50
E $ 1.00

(c) For any real numbers A and B, A * B is defined as 2B - A. What is the value of 4 * 3?
A 2
B 3 * 4
C 5
D 12
E None of the above

Essay. This test attempted to assess pupils' ability to communicate correctly and fluently in writing. They were asked to write an essay that reflected the four following characteristics:

(1) interest and liveliness;

(2) correct grammar, spelling and punctuation;

(3) a clear and fluent writing style; and

(4) appropriate organization.

Pupils were asked to write three or four paragraphs (approximately 200 words) on any one of six possible topics offered. The six topics were evenly distributed over three types of essay - expository, imaginative and emotional. The time limit to complete the essay was 25 minutes. A copy of the Essay Test can be seen in Appendix H.

The marking scheme for the essays (see Appendix Hi) focussed on four aspects: (1) Mechanics (i.e. length, grammar, paragraphing, opening and conclusion); (2) Content (i.e. the colourfulness and depth of vocabulary); (3) Style (i.e. the use of appropriate language, and originality in both ideas and expression); and (4) Organisation (i.e. logical and effective use of facts and ideas). Each of these components contributed an equal weighting to the total score.

General Scholastic Aptitude (GSAT). This test attempted to measure how well pupils think in relation to general school-related tasks. Although closer to a "true" syllabus-free test than the developed abilities tests, GSAT was also largely independent of any school-related study, as well as lacking face and content validity with respect to specific subject disciplines. GSAT most closely compares with Otis and TOSCA, and probably the best definition of what it measures relates to Spearman's definition of intelligence (i.e. "g").

The test was developed in part from the work of Gilmore (1979, pp 71-2) and consisted of different kinds of word, number

and reasoning problems which were designed to measure the ability to identify relationships and to apply logical reasoning. The inclusion of word knowledge in an aptitude test may seem somewhat contradictory. However, it has been argued (Anderson and Freebody, 1981, p 81-5) that vocabulary is, in fact, a good indicator of verbal aptitude, that is, the "mental agility" required for the successful comprehension of word meanings. Thus, pupils who score higher on word knowledge tend also to have a higher degree of verbal aptitude.

The test was divided into four sections, beginning with:

SECTION 1 - Word Knowledge (e.g. word meanings, opposites, etc).

Example: Oblivious
A painless
B static
C eternal
D dead
E unmindful

SECTION 2 - Relationships Between Words (e.g. verbal analogies, odd word out, meaning of proverbs, etc).

Example: Light is to dark as pleasure is to
A picnic
B day
C pain
D night
E boat

SECTION 3 - Relationship Between Elements (e.g. numerical, alphabetical and spatial series, etc).

Example: What is the next number in the series?
1 3 5 7 9 11
A 12
B 13
C 14
D 15
E None of the above.

SECTION 4 - Verbal (or logical) Reasoning.

Example: John runs faster than Fred. Fred runs faster than Ian. Tony and Sid are slower than Ian. Who runs the fastest?

A John
B Fred
C Ian
D Tony
E Sid

There were a total of 25 questions to be answered within a 25 minute time limit. A copy of the GSAT can be seen in Appendix I.

Rationale Underlying the Item Types

A number of advantages associated with objective item formats were also identified as important in the development of a moderation test. Easy administration, fast accurate marking, wide content coverage, precise problem posed, easier specification of test content and disposition to pretesting - all lend themselves well to the construction of a highly objective, valid and reliable instrument for group discrimination.

The most popular and technically efficient of the objective item types is the multiple-choice format, especially with professionally produced tests. It has become the "standard bearer" by which all other item types are compared and evaluated.

However, the problems of "test-wiseness" with pupils who are prepared "to take chances guessing" may give some an unfair advantage because of the popularity of the multiple-choice question. As well, different subject requirements (e.g. English versus mathematics) suggest the need for investigation of alternative item formats, such as the open-ended form and cloze procedure.

Survey evidence from Elley (in progress), and a suggestion for further research by Hulbert (1978) have shown that many teachers are not especially keen on the multiple-choice item, thus confirming the need for investigation in this area.

Description of the Test Items

Multiple-Choice. This item was used exclusively in one set of parallel tests, thus allowing a direct comparison with the alternative item tests. The multiple-choice format was used in tests of Vocabulary and Comprehension (i.e. English, science and social studies), Mathematics and GSAT. The five option format was selected to ensure maximum discrimination in conjunction with the relative shortness of the tests (See evidence, pp 52-3). Multiple choice forms of the tests can be seen in Appendices B, D and F.

Open-Ended (or Short Answer). This item was used in the alternative parallel tests as a comparison with the multiple-choice tests. The open-ended format was employed exclusively in the Vocabulary and Mathematics tests, and partly in the Comprehension Test to assess the poetry, graph and map questions. Open-ended forms of the tests can be seen in Appendices C and G. Two examples are presented below:

EXAMPLE 1: The tangi lasted for several days.

ANSWER IS WRITTEN HERE...

EXAMPLE 2: If 4n + 7 = 5, then n is equal to

..AND HERE.

Cloze Procedure. This item format was also used in the alternative parallel tests as a comparison with the multiple-choice tests. The cloze procedure was used to assess the prose passages in the Comprehension Test. Every 10th word was omitted from the passage and replaced by a standard length gap, after leaving the first paragraph unaltered (rather than the usual first sentence, due to the length of the passages). The cloze test can be seen in Appendix E.

An analysis of different cloze marking procedures was also conducted, comparing exact (or verbatim) replacement with synonym replacement. The former allows for absolutely no divergence from the original text (disregarding mis-spellings), and although this may at times appear a very harsh method, it does have the major advantage of simplifying the marking and making it a totally objective process. Synonym replacement, on the other hand, gives credit to non-verbatim responses which are otherwise syntactically and/or semantically correct. There is, however, likely to be a corresponding decrease in the objectiveness and reliability of this type of marking process. An example of the two marking procedures follows:

Text: "It requires the recipient to locate in the rēsumē
__1__ qualification he has asked for. It may, if
the __2__ is running more than one ad..."

Exact Replacement: 1. the 2. firm

Synonym Replacements:

1. which; each; every; any (the)
2. company; employer; business; department (firm)

DATA COLLECTION

The tests were administered over a three week period from mid October to early November, 1986.

The final timetabling of the test sessions was dependent on the individual schools concerned, with the only stipulation being that the total testing time (140 minutes) be divided into either two or three sessions to avoid a daunting test-battery being presented to pupils. Two of the schools chose a combination of two-hour and one-hour sessions while the other two opted for three one-hour sessions.

A senior teacher acted as a liaison between the schools and the researcher. At least one visit was made to each of the schools by the researcher prior to the commencement of the testing. In two of the schools the tests were administered by the individual class teachers under the guidance and general supervision of the researcher, while in the other two, the researcher and several colleagues conducted the test administration (see Directions for Supervisors, Appendix K). Every effort was made to standardize the administration of the tests and the conditions under which pupils took the tests. For each session there were four different test forms (multiple-choice A and B; open-ended/cloze A and B) which were distributed alternatively in a random fashion around the classroom.

Computer listings of the School Certificate Examination marks were obtained from each school as they became available

to the researcher during February, 1987. These marks were used as the criterion for judging the predictive validity of the reference tests.

Table 3.3 presents the relevant details of all the reference tests, number of items and administration time.

TABLE 3.3

General Details for all Reference Tests

Test	N of Forms	N of Items per Test	N of Items Tested	Time for Administration
English Vocabulary	4	12	48	5 mins
Science Vocabulary	4	12	48	5 mins
Social St. Vocab.	4	12	48	5 mins
English Comprehension	4	17 (49)*	34 (98)	16.7 mins
Science Comprehension	4	17 (44)	34 (88)	16.7 mins
Social St. Comp.	4	17 (50)	34 (100)	16.7 mins
Mathematics	4	25	100	25 mins
GSAT	1	25	25	25 mins
Essay	1	6	6	25 mins
Total:	-	143 (235)	377 (561)	140 mins

* The figures in brackets represent the number of items in the cloze tests. For all other tests, the question format makes no difference to the number of items in each test.

DATA ANALYSIS

A number of preliminary analyses were carried out as a check on the validity of the data, before considering the major investigation of the study. The preliminary analyses are listed as follows:

(1) a check on the completeness of the data obtained;

(2) a check of the marking patterns from the multiple-choice tests. This indicated the need for an item analysis to be conducted on a limited number of suspect items due to unexpected outcomes caused by either a faulty distractor or that proved just too difficult;

(3) descriptive statistics (mean, standard deviation, etc.) were generated for all reference tests based on their raw scores, and for the following School Certificate Examination marks: English, Mathematics, Science[1], Geography, History and Economics. Any one of these last three were used as the criterion for the Social Studies reference test. If a pupil had done two or all three of them, then an average mark was calculated;

(4) to allow the conversion of the reference test raw scores into standard scores, the assumption of a completely random administration of the parallel forms for each test (Vocabulary, Comprehension and Mathematics), was checked by analysing four sub-samples who each sat one of four Vocabulary Test forms. The results from a "common" test (GSAT) taken by the total sample, were then compared across the respective sub-samples;

(5) the reference test raw scores were converted to T-scores, with a mean of 50 and a standard deviation of 10, to enable a direct comparison across the different test forms; and

(6) a check on the reliability of the tests was made using the split-half r method, and on the reliability of the essay marking through an inter-rater comparison.

The first stage of the major analysis involved the generation of a series of Pearson's product-moment correlation matrices

---

[1]For one high ability class 7 pupils were doing Chemistry and Physics instead of general Science. In these cases, an average mark was calculated.

across all likely variables of interest (i.e. moderation test scores, School Certificate Examination marks, item formats and sex).

The second stage focussed on the prediction of class parameters (mean and standard deviation) for the School Certificate Examination marks from the performance of the pupils on the moderation tests. This was completed using a combination of stepwise and forced entry multiple-regression analyses. Checks for the degree of shrinkage of the multiple correlations were also made.

The computer analyses were run using statistical packages from the SPSS series (SPSSX, 1983).

CHAPTER IV

RESULTS

COMPLETENESS OF DATA

Invariably, with this type of empirical research, it is impossible to collect a full set of data. The reasons for this were as follows:

(1) administration of the tests over two or three days meant that pupils absent on any one day would have incomplete data;

(2) poor co-ordination between the liaison teacher and the classroom teachers resulted in minor confusion at two schools. Firstly, about 20 pupils missed out on a major testing session when they were directed to the wrong classroom; and secondly - despite fairly implicit instructions - two teachers administering the Comprehension Test became confused by the different forms, and instead gave out more multiple-choice than cloze tests; and

(3) one complete class was lost - in addition to individual pupils from other classes - when a review of the School Certificate results showed these pupils had entered for only one or two School Certificate Examinations with most of these being

of a practical nature (e.g. Woodwork, Typing), and therefore not providing a suitable criterion measure relevant to the study's aims.

Thus, the median percentage of pupils across all schools who were included in the final computer analyses was 86% (range 81% to 90%) of the original experimental sample (see Table 3.1). Details of the final number of tests completed and included for further data analysis are presented in Table 4.1.

TABLE 4.1

The Number of Completed Reference Tests, by School, Included in the Final Data Analysis.

	School	Vocabulary	Comprehension	Maths	Essay	GSAT
	1	92	88	91	92	92
	2	102	96	104	90	90
	3	92	89	92	93	93
	4	83	87	84	83	83
Total:	4	369	360	371	358	358

Estimation of Scores for Incomplete Tests

After the first administration of the Comprehension Tests it became apparent that the open-ended/cloze forms were too long for the majority of pupils to finish in the time allocated. The tests were subsequently reduced in length and administered to the rest of the sample without problem. However, to avoid

the partial loss of data from the first administration, individual scores on the open-ended/cloze forms were estimated for any incomplete tests using the regression equation:

$$\tilde{Y} = a + bx$$

where $\tilde{Y}$ is the total score (or criterion) on the Comprehension Test to be predicted;
a is the intercept on the Y-axis;
b is the slope of the regression line; and
x is the sub-test score (or predictor variable) from the incomplete Comprehension Test.

This process was especially important since the sample size for such an extensive study was relatively modest. The same regression equation was also used to predict missing scores for the Vocabulary and Mathematics Tests.

For the remainder of this chapter, results have been presented as separate entities for the Vocabulary and Comprehension Tests. This was carried out mainly for research purposes. However, in terms of policy implications, it is only the combined Vocabulary/ Comprehension format that is under serious consideration as a moderating test.

DESCRIPTIVE AND RELIABILITY ANALYSES OF THE TEST DATA

This section is concerned with the presentation of the technical characteristics of the reference tests and School Certificate Examination marks. The characteristics to be reported include the mean, standard deviation, number of cases, number of items and an estimate of test reliability, based on the split-half

correlation corrected for length by the Spearman-Brown formula. The reliability analysis was conducted on Form A only of each test. Since the different forms had in general comparable means and standard deviations, it was considered unlikely that the Form B estimates would differ greatly. In addition, it was felt that the reliability analysis was very much a secondary aspect of the study (although still an important factor) and did not warrant further time being spent on it.

Vocabulary (English, Science and Social Studies)

The first set of results come from the English, Science and Social Studies Vocabulary Tests. The different test formats by question type (i.e. the multiple-choice and open-ended tests) have been combined to illustrate the technical characteristics of the respective Vocabulary Tests overall. These have been reported below in Table 4.2.

TABLE 4.2

Descriptive Statistics of the English, Science and Social Studies Vocabulary Tests Across Item Types.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	6.16	5.66	6.25	6.26	5.98	5.42
S.D.	2.48	2.66	2.54	2.28	2.43	2.43
N of Cases	188	181	184	179	174	178
N of Items	12	12	12	12	12	12
Split-Half r	0.70	-	0.62	-	0.67	-

Generally, the means and standard deviations for Forms A and B of each Vocabulary Test were quite comparable, the largest mean difference (0.56) occurring in Social Studies. The reliability estimates (ranging from 0.62 for Science to 0.70 for English), while not exceptionally high, are in fact most satisfactory for a short 12 item test.

Tables 4.3 and 4.4 present in succession below the results of the English, Science and Social Studies Vocabulary Tests as separate multiple-choice and open-ended analyses respectively.

TABLE 4.3

Descriptive Statistics of the English, Science and Social Studies Multiple-Choice Vocabulary Tests

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	7.20	6.40	6.31	6.99	5.93	5.87
S.D.	2.43	3.07	2.67	2.38	2.52	2.80
N of Cases	98	98	97	97	97	97
N of Items	12	12	12	12	12	12
Split-Half r	0.56	-	0.66	-	0.60	-

TABLE 4.4

Descriptive Statistics of the English, Science and Social Studies Open-Ended Vocabulary Tests

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	5.02	4.79	6.17	5.39	5.36	4.88
S.D.	2.52	2.25	2.41	2.17	2.35	2.07
N of Cases	90	83	87	82	87	81
N of Items	12	12	12	12	12	12
Split-Half r	0.80	-	0.65	-	0.74	-

Once again, a look at the means and standard deviations reveals that they are reasonably similar across the alternate forms in each subject test for both the multiple-choice and open-ended formats. The mean differences are slightly larger (ranging from 0.06 to 0.80 for the multiple-choice and 0.23 to 0.78 for the open-ended) than those observed for the combined Vocabulary results in Table 4.2. This was probably due to the smaller number of cases involved in the two sub-analyses (median N's of 97 and 85) compared with the combined analysis (median N of 180). A comparison of the multiple-choice and open-ended results shows the former having produced small, but consistently higher mean scores than the latter.

Of more significance is the difference in the reliability estimates for the two sub-analyses. The split-half correlations for the multiple-choice tests ranged from 0.56 (English) to 0.66 (Science), while those for the open-ended tests ranged from 0.65 (Science) to 0.80 (English). However, this major improvement occurred in only two of the three subject tests, with Science seemingly unaffected by the change in item format.

Comprehension (English, Science and Social Studies)

The second set of results concerns the English, Science and Social Studies Comprehension Tests. Unlike the Vocabulary, the multiple-choice and cloze versions of the Comprehension Test differ greatly in the number of items in each (17 vs 48). Therefore, to conduct a combined analysis across item types, although possible, would be mathematically tedious and since all the tests were converted to standard scores for the subsequent validity analysis, this extra step was considered unnecessary. Thus, only the results of separate sub-analyses for the multiple-choice and cloze Comprehension Tests have been presented below in Tables 4.5 and 4.6 respectively.

TABLE 4.5

Descriptive Statistics of the English, Science and Social Studies Multiple-Choice Comprehension Tests.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	9.79	9.79	8.73	9.27	7.10	6.13
S.D.	3.27	3.27	3.70	3.08	3.28	3.16
N of Cases	100	99	100	98	100	98
N of Items	17	17	17	16*	17	16*
Split-Half r	0.66	-	0.74	-	0.72	-

*The reduction in the number of items was necessitated when an item analysis check revealed two items with very poor discrimination qualities.

TABLE 4.6

Descriptive Statistics of the English, Science and Social Studies Cloze Comprehension Tests.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	22.98	17.33	18.60	14.06	12.15	16.90
S.D.	5.12	5.66	5.80	4.59	3.88	5.31
N of Cases	83	78	77	73	78	72
N of Items	49	48	44	44	50	51
Split-Half r	0.89	-	0.91	-	0.79	-

In the multiple-choice format of the Comprehension Test there was some minor variation in the means and standard deviations of the alternate forms for Social Studies and Science, while the English Tests equated exactly. The reliability estimates ranged from 0.66 (English) to 0.75 (Science); again, these seem quite satisfactory for a relatively short 17 item test. The cloze format of the Comprehension Test displayed a greater degree of variation in the means and standard deviations for all three subjects, with the largest mean difference of 5.65 occurring in English. However, the reliability estimates were significantly higher in comparison to the multiple-choice tests, ranging from 0.79 (Social Studies) to 0.91 (Science). This increase was due largely to the increased number of items utilized in the cloze procedure. Thus, the cloze is clearly a more efficient item format, in that it produces a measure of comprehension that has higher reliability per unit of testing time.

A secondary analysis of the cloze Comprehension Test incorporated synonym replacement marking (as compared with exact or verbatim replacement). The technical data related to this marking procedure has been reported below in Table 4.7.

TABLE 4.7

Descriptive Statistics* of the English, Science and Social Studies Cloze Comprehension Tests, Using Synonym Replacement Marking.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	33.87	26.99	25.70	21.91	21.93	25.38
S.D.	6.06	7.53	6.79	5.77	5.83	8.76
Split-Half r	0.94	-	0.93	-	0.91	-

*The N of Cases and Items were excluded since these figures remain unchanged from those reported in Table 4.6.

The pattern of variation in the means and standard deviations remained the same under the synonym marking scheme, although of course, both parameters showed consistent increases for all subjects. The reliability estimates improved slightly for both English and Science, while that for Social Studies jumped markedly from 0.79 to 0.91.

Vocabulary and Comprehension Combined

The third set of results focusses on the Vocabulary and Comprehension Tests as a combined analysis, to assess these two components as a single reference test for English, Science and Social Studies. The results have been presented as separate item type analyses in Tables 4.8 and 4.9.

TABLE 4.8

Descriptive Statistics for the English, Science and Social Studies Multiple-Choice Vocabulary/Comprehension Tests.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	16.99	16.19	15.04	16.26	13.03	12.00
S.D.	5.70	6.34	6.37	5.46	5.80	5.96
N of Cases	98	98	97	97	97	97
N of Items	29	29	29	28*	29	28*
Split-Half r	0.76	-	0.74	-	0.81	-

*Items excluded due to poor discrimination qualities.

TABLE 4.9

Descriptive Statistics for the English, Science and Social Studies Open-Ended Vocabulary/Cloze Comprehension Tests.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	28.00	22.12	24.77	19.45	17.51	21.78
S.D.	7.64	7.91	8.21	6.76	6.23	7.38
N of Cases	83	78	77	73	78	72
N of Items	61	60	56	56	62	63
Split-Half r	0.97	-	0.94	-	0.84	-

The multiple-choice format of the combined Vocabulary/ Comprehension Tests, again produced means and standard deviations that were very similar across the alternate A and B forms in all three subjects. The increased length of the combined test (29 items) resulted in an improvement on the estimates for the single Vocabulary and Comprehension components in English and Social Studies (see Tables 4.3 and 4.4). The split-half correlations increased to 0.76 for English and 0.81 for Social Studies. However, the combined estimate for Science of 0.74 was the same as that reported for the Science Comprehension Tests alone (see Table 4.5).

Consistent with the earlier results, there was some variation in the mean and standard deviation scores across alternate forms for the combined open-ended Vocabulary/cloze Comprehension Tests. The increase in length of the combined tests (44 to 51 items) improved the reliability estimates by 0.10 for English (to 0.97), 0.03 for Science (to 0.94) and 0.05 for Social Studies (to 0.84) on the higher single test estimates of Comprehension alone.

The higher reliability estimates generated from the cloze Comprehension format was a direct outcome of the greater number of items used. An equivalent number of items in a multiple-choice format would no doubt produce similarly high reliability estimates, but would take more time for pupils to complete.

Would the use of synonym replacement marking with the cloze Comprehension Tests improve the reliability any further?

The results of this analysis are reported below in Table 4.10.

TABLE 4.10

Descriptive Statistics* of the English, Science and Social Studies Open-Ended Vocabulary/Cloze Comprehension Tests Using Synonym Replacement Marking.

Subject	English		Science		Social Studies	
Form	A	B	A	B	A	B
M	38.89	31.78	31.87	27.30	27.29	30.26
S.D.	8.58	9.78	9.20	7.94	8.18	10.83
Split-Half r	0.99	-	0.96	-	0.95	-

*The N of Cases and Items were excluded since these figures remain unchanged from those reported in Table 4.9.

Consistent with the improvement noted previously, the use of synonym replacement marking increased the reliability estimates marginally for English and Science (+0.02), but rather more significantly for Social Studies from 0.84 to 0.95 (+0.11). The large gain in reliability for the Social Studies Tests, is probably a reflection of the lower mean scores in the cloze using exact replacement, this being supplemented to a greater degree by the use of synonym replacement than were the English and Science tests. In other words, the passages used in the Social Studies Test were slightly too difficult for Form 5 pupils, when exact replacements were required.

Mathematics

The fourth set of results was concerned with the Mathematics Tests in both an overall analysis across item types and as sub-analyses of the multiple-choice and open-ended formats. The first of these has been presented below in Table 4.11.

TABLE 4.11

Descriptive Statistics of the Mathematics Tests Across Item Types.

Form	M	S.D.	N of Cases	N of Items	Split-Half r
A	12.04	4.96	190	25	0.88
B	15.64	5.48	181	25	-

A difference of 3.60 between the two means would suggest that Form A of the Mathematics test was slightly more difficult than Form B. The standard deviations were fairly similar, while the reliability estimate of 0.88 was excellent for a test of moderate length (25 items).

The results from the two item type sub-analyses have been reported in succession below in Tables 4.12 and 4.13.

TABLE 4.12

Descriptive Statistics of the Multiple-Choice Mathematics Tests.

Form	M	S.D.	N of Cases	N of Items	Split-Half r
A	12.46	4.85	97	25	0.86
B	16.22	5.61	95	25	-

TABLE 4.13

Descriptive Statistics of the Open-Ended Mathematics Tests.

Form	M	S.D.	N of Cases	N of Items	Split-Half r
A	11.60	5.06	93	25	0.88
B	14.99	5.35	86	25	-

As with the overall results, there was a moderate difference (3.50) between the means of the alternate forms for both the multiple-choice and open-ended versions of the test. The standard deviations were reasonably comparable between forms in each case. Similarly, the respective item formats made no significant difference to the estimates of each test's reliability. The reliability

correlations of 0.86 (multiple-choice) and 0.88 (open-ended) were excellent for such relatively short (25 item) tests.

Essay

The fifth set of results comes from the Essay Test. These were marked employing four equally weighted sub-components to generate a total score out of 20. These sub- and total-analyses have been reported in Table 4.14 below.

TABLE 4.14

Descriptive Statistics of the Mechanics, Content, Style, Organisation and Total Essay Scores.

	Mechanics	Content	Style	Organis-ation	Total
M	3.54	3.18	3.12	2.90	12.74
S.D.	1.09	0.60	0.78	0.74	3.21
Max. Possible Score	5	5	5	5	20
N of Cases	358	358	358	358	358
N of Items	1	1	1	1	1

All the essays were marked initially by the researcher. However, because ofthe known variation that exists due to the subjective nature of essay marking, an inter-rater reliability check was made employing a fellow graduate student. The secondary marking

was conducted on a random one-in-two sample (N = 191) of the total number of essays in the study. All marks were removed from the sample of essays in readiness for the remark and the same marking schedule was used by the check-marker. Pearson's product-moment correlation was used to analyse the two sets of results which were based on Total scores. The results of inter-rater reliability analysis have been reported below in Table 4.15.

TABLE 4.15

Means, Standard Deviations and Correlation of the Inter-Rater Reliability Estimate for the Essay Marking.

	Rater 1		Rater 2		
	M	S.D.	M	S.D.	
N = 191	12.87	2.61	13.30	2.55	$r = 0.64^*$

*Statistically significant at $p < 0.01$.

The reliability index of 0.64 was in line with other estimates of reliability under similar conditions (e.g. Elley, Barham, Lamb and Wyllie, 1979).

General Scholastic Aptitude Test (GSAT)

The technical characteristics of the GSAT[1], the final moderating test in this section, have been presented below in Table 4.16.

---

[1]There is only one form of GSAT, and therefore, no comparison between item types.

TABLE 4.16

Descriptive Statistics of the General Scholastic Aptitude Test.

	M	S.D.	Nof Cases	Nof Items	Split-Half r
GSAT	16.61	4.03	358	25	0.78*

*This figure was based on a random one-in-two sample, N = 179

A large number of higher scores and a fairly high mean indicate a possible ceiling effect to some extent, due mainly to a combination of the relative ease of the items, and the test's length per unit of time. As with the other tests, a reliability estimate of 0.78 was very good in relation to the modest length of the test (25 items).

School Certificate Examinations

By reporting the descriptive statistics of the School Certificate Examinations an indication can be gained as to how well the performance of the sample equates with that of the nations's fifth formers as a whole. However, since there were no mean or median estimates available for the national School Certificate Examination results, the best option was to calculate percentage pass rates to provide a suitable comparison. Obviously, no estimate of the reliability can be obtained for the examinations. However, the remaining results of the descriptive analysis and the percentage pass rates have been presented below in Table 4.17.

TABLE 4.17

Descriptive Statistics and Percentage Pass Rate of the School Certificate Examinations.

Subject	M	S.D.	N	Percentage Pass Rate	
				Nat. Pop.*	Sample
English	53.58	16.08	393	59.0%	60.6%
Mathematics	53.62	15.79	374	63.9%	69.0%
Science	54.91	16.52	328	63.1%	64.9%
Social Studies	56.06	16.01	269	62.6%	65.1%
Total	224.73	55.74	378	-	-

*School Certificate Examination Statistics 1985, Department of Education, Wellington, New Zealand, 1986. (Table 14, p 20)

A comparison of the respective School Certificate Examination pass rates for the national population and experimental sample were clearly very similar. The largest percentage difference (5.1) occurred in Mathematics. The sample results were consistently higher than the national figures, although not significantly. It must be concluded, based on the School Certificate percentage pass rates, that the performance of the sample of fifth formers in the current study was approximately equal to that of the nation's fifth formers as a whole.

CONVERSION OF RAW SCORES TO STANDARD SCORES

To enable a direct comparison across the alternate test forms and across the different item formats, all raw scores from the reference tests were converted to a standard score with a specified mean and standard deviation. In addition to allowing a combined analysis of the multiple-choice and open-ended/cloze tests, it also countered any problem with differences in the difficulty level of the alternate forms. For example, Form B of the Mathematics Test was more difficult than Form A, while there was more variation amongst the means and standard deviations of the open-ended/cloze tests.

As a particularly convenient scale, it was decided to convert the raw scores to T-scores with a mean of 50 and a standard deviation of 10. The basic standard score formula used in the conversion was:

$$50 + 10 \frac{(X-M)}{S.D.}$$

In converting to standard scores it is being assumed that there was a completely random administration of the parallel forms of each test. This assumption was tested by analysing one of the multiple-form tests (Vocabulary) against the results of a common test sat by the total sample. The results of this analysis are reported below in Table 4.18.

TABLE 4.18

Mean and Standard Deviation of the GSAT Scores for the Sub-Samples Administered the Four Multiple-Forms of the Vocabulary Test.

Item Type	Multiple-Choice		Open-Ended	
Form	A	B	A	B
M	49.92	50.54	49.51	50.62
S.D.	8.58	9.93	10.25	10.96
N	93	87	83	81

Clearly, there was no significant difference in the parameters of the GSAT scores between the respective sub-samples, therefore, the assumption of a completely random administration appears to hold true. That is, the range of pupil abilities given each form of the test was approximately equal.

SEX DIFFERENCE IN THE REFERENCE TESTS AND SCHOOL CERTIFICATE EXAMINATIONS

One further characteristic of interest was the relative performances of the two sexes on the reference tests and School Certificate Examinations. This analysis provided an indication of any possible or unexpected sex bias among the tests. Descriptive

statistics were generated as separate female and male analyses, and then t-tests were conducted to assess the statistical significance of the variation in mean scores. The results have been reported below in Table 4.19.

TABLE 4.19

Descriptive Statistics and t-Tests for Males and Females of Reference Tests and School Certificate Examinations.

Test	Females			Males			t-Test
	M	S.D.	N	M	S.D.	N	
Reference:							
Engl Vocab + Comp	49.75	10.03	181	50.20	9.93	178	0.43, n.s.
Scie " + "	49.07	10.03	176	50.90	9.77	171	1.71, $p<.05$
Sost " + "	49.21	9.92	176	50.73	10.03	171	1.42, n.s.
Mathematics	47.97	9.84	181	52.01	9.77	190	3.96, $p<.01$
Essay	52.54	8.98	175	47.57	10.32	183	4.83, $p<.01$
GSAT	49.03	9.54	175	50.93	10.36	183	1.85, $p<.05$
School Certificate:							
English	56.88	16.08	195	50.33	15.44	198	4.12, $p<.01$
Mathematics	51.84	15.70	182	55.30	15.74	192	2.12, $p<.05$
Science	51.83	16.38	157	57.74	16.18	171	3.27, $p<.01$
Social Studies	54.01	16.99	143	58.39	14.55	126	2.27, $p<.05$
Total	223.80	56.66	185	225.62	54.98	193	0.64, n.s.

Of the reference tests, differences were found in the mean scores favouring males for the Mathematics, Science and GSAT Tests, while females clearly did better on the Essay Test. There was no significant difference between scores for the English and Social Studies Tests. In comparison, the School Certificate Examinations saw males performing significantly better in Mathematics, Science and Social Studies, while females did similarly better in English. However, despite these variations, both sexes performed equally well in terms of their School Certificate Total scores.

A comparison of the tests and the examinations showed the sex differences to be consistent for Mathematics and Science, but not for Social Studies or English. Although females performed favourably in School Certificate English and the Essay Test, this bias did not carry over into the English Vocabulary/Comprehension Test, at least not sufficiently to register as significant.

PREDICTIVE VALIDITY OF THE REFERENCE TESTS

This section of the results focusses on the prediction of the School Certificate Examination marks using the reference or moderating tests. In so doing, and as the crux of the study, the performance of the reference tests can be examined as suitable predictors of the School Certificate Examinations. For moderation and policy purposes, the important unit of analysis is the class, not the individual. However, the results based on individual scores

have also been presented, mainly to enable a comparison with earlier studies which only focussed at this level of analysis, and for reasons of interest as well.

The presentation of the validity results is in three main parts. Firstly, the correlations between the reference tests and the School Certificate Examination marks based on individual pupils' scores; secondly, the prediction of the School Certificate results based on class (or group) performances measured by mean and standard deviation estimates; and thirdly, the use of multiple-regression analyses to find the best weighting of a combination of reference tests in the prediction of the class-based performances of the School Certificate Examinations. In addition,the results of the respective sub-analyses for the multiple-choice and open-ended test formats will be presented also.

It should be noted that all correlation coefficients in the validity analysis were derived using Pearson's product-moment formula.

Correlations Between the Reference Tests and School Certificate Marks Based on Individual Scores

The first set of results to be reported is a table of correlations presenting the results of a combined analysis across the different question formats based on pupils' individual scores only. The results of this analysis can be seen below in Table 4.20.

TABLE 4.20

Correlations Between the Reference Tests and School Certificate Examination Marks, Based on Pupils' Individual Scores.

Reference Test	School Certificate				
	English	Mathematics	Science	Social Studies*	Total+
Vocabulary:					
English	(.63)	.55	.61	.64	.66
Science	.58	.60	(.71)	.63	.67
Social Studies	.61	.57	.68	(.68)	.67
Total++	.69	.65	.76	.75	(.76)
Comprehension:					
English	(.60)	.46	.53	.60	.61
Science	.57	.58	(.63)	.60	.65
Social Studies	.60	.59	.63	(.64)	.68
Total++	.66	.62	.76	.75	(.76)
Vocab + Comp:					
English	(.69)	.58	.66	.71	.72
Science	.67	.67	(.76)	.70	.76
Social Studies	.68	.65	.73	(.74)	.75
Total++	.73	.69	.78	.77	(.81)
Mathematics	.53	(.79)	.72	.66	.73
Essay	(.59)	.19**	.26	.32	.41
GSAT	(.59)	(.61)	(.65)	(.64)	(.68)

*Combination of Geography, History and/or Economic Studies

+English and best 3 other School Certificate Examinations

++English, Science and Social Studies Combined

**All r's statistically significant at $p<0.01$, except where marked with a double asterisk which is n.s.

()Brackets denote the most important relationships between corresponding subjects

Ignoring some small correlations obtained with the Essay Test, the others ranged from 0.41 to 0.81. The majority of the coefficients (82.4 percent) were spread from 0.41 to 0.59. In nearly all cases, the highest correlation with the respective School Certificate subjects was the corresponding subject test (i.e. the English reference tests had the highest correlation with School Certificate English, the Mathematics reference test had the highest correlation with School Certificate Mathematics, and so on.) The combining of the Vocabulary and Comprehension Tests resulted in higher correlations than for the individual tests, while the Essay Test correlated with School Certificate English as well as the English Comprehension Test did.

The combined total tests (Englsh, Science and Social Studies) for Vocabulary, Comprehension and Vocabulary/Comprehension produced moderately high correlations (median r's 0.75 or greater) with all the School Certificate Subjects. This would seem to indicate that pupils performing well in one subject, also tend to do well in other subjects. The Mathematics and combined Vocabulary/Comprehension Tests also correlated quite highly (median r's 0.69 to 0.73).

The results of the two sub-analyses for the multiple-choice and open-ended formats of the reference tests have been reported in Tables 4.21 and 4.22.

TABLE 4.21

Correlation Between the Multiple-Choice Reference Tests and School Certificate Marks, Based on Pupils' Individual Scores.

Reference Test	School Certificate				
	English	Mathematics	Science	Social Studies*	Total+
Vocabulary:					
English	(.65)	.58	.63	.66	.68
Science	.55	.58	(.70)	.62	.64
Social Studies	.59	.57	.68	(.70)	.65
Comprehension:					
English	(.57)	.43	.50	.60	.59
Science	.59	.64	(.68)	.65	.68
Social Studies	.62	.62	.66	(.72)	.70
Mathematics	.52	(.79)	.76	.67	.73

All r's statistically significant at $p < 0.01$.

TABLE 4.22

Correlations Between the Open-Ended/Cloze Reference Tests and School Certificate Marks, Based on Pupils' Individual Scores.

Reference Test	School Certificate				
	English	Mathematics	Science	Social Studies*	Total+
Vocabulary:					
English	(.61)	.54	.59	.62	.65
Science	.61	.62	(.73)	.65	.72
Social Studies	.64	.57	.68	(.64)	.69
Comprehension:					
English	(.65)	.51	.59	.61	.66
Science	.56	.51	(.60)	.55	.63
Social Studies	.59	.56	.61	(.54)	.67
Comprehension (SR)*:					
English	(.67)	.53	.63	.57	.68
Science	.59	.54	(.59)	.55	.65
Social Studies	.61	.57	.62	(.56)	.67
Mathematics	.53	(.80)	.70	.66	.75

*Cloze tests marked using Synonym Replacement.

All r's statistically significant at $p < 0.01$.

A comparison of the two sets of results revealed little or no significant difference in the correlations obtained, whether using the multiple-choice or open-ended/cloze question format. The correlations were quite stable when comparing the respective Vocabulary Tests, Mathematics Tests and for the multiple-choice Comprehension Tests with the cloze Comprehension Tests using synonym replacement marking. The cloze Comprehension Test, using exact replacement marking, produced correlations that were generally slightly lower than the other two comprehension tests, however the differences were fairly minimal.

Prediction of School Certificate Class Parameters: Mean and Standard Deviation

The prediction of class performances using mean and standard deviation estimates, as mentioned earlier, represents the key problem in moderation. The prediction of class means provides a measure of the relative ability level of each class, and therefore, is the most important of the two parameters. The prediction of class standard deviations, although less crucial in the context of moderation, provides a useful guide to the range or spread of abilities in each class. These analyses were based on a sample of 18 classes.

As before, the results have been presented as a combined analysis and then as separate sub-analyses by item type. The results of the mean and standard deviation analyses will also be presented in separate tables. The first of these results,the

prediction of class means and standard deviations across item formats, have been reported below in Tables 4.23 and 4.24.

TABLE 4.23

Prediction of School Certificate Class Means Across Item Types (N=18)

Reference Test	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocabulary:					
English	(.87)	.81	.84	.80	.89
Science	.77	.79	(.83)	.79	.85
Social Studies	.78	.80	.92	(.78)	.88
Total	.87	.84	.92	.83	(.92)
Comprehension:					
English	(.76)	.84	.87	.72	.83
Science	.81	.88	(.88)	.76	.87
Social Studies	.84	.90	.88	(.77)	.90
Total	.84	.91	.89	.77	(.90)
Mathematics	.79	(.97)	.91	.86	.94
Essay	(.62)	.40	.50	.32*	.48
GSAT	(.86)	(.90)	(.87)	(.79)	(.89)

*All r's statistically significant at $p < 0.05$, except where marked with an asterisk.

TABLE 4.24

Prediction of School Certificate Class Standard Deviations Across Item Types (N = 18)

Reference Test	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocabulary:					
English	(.29)*	.17*	.16*	.45	.24*
Science	.31*	.20*	(.28)*	.43	.35*
Social Studies	.06*	.09*	-.07*	(.31)*	.09*
Total	.38	.23*	.33*	.55	(.37)*
Comprehension:					
English	(.46)	.47	.22*	.39	.47
Science	.37*	.56	(.44)	.59	.53
Social Studies	.28*	.46	.23*	(.43)	.34*
Total	.51	.61	.47	.61	(.62)
Mathematics	.70	(.57)	.43	.31*	.64
Essay	(.32)*	.20*	.48	.36*	.41
GSAT	(.58)	(.70)	(.61)	(.67)	(.70)

*Are statistically not significant; all other r's significant at $p < 0.05$.

The prediction of the class means produced a series of relatively high correlations (excluding the Essay Test) ranging from 0.72 to 0.97. A median correlation of 0.83 was recorded across the four Vocabulary Tests, 0.87 for the Comprehension Tests, 0.90 for the combined Vocabulary/Comprehension Tests, 0.91 for the Mathematics Test and 0.87 for the GSAT. The Essay Test correlated moderately well with School Certificate English at 0.62.

The final point of interest from Table 4.23 was that the corresponding subject tests (i.e. English reference tests predicting School Certificate English, Science reference tests predicting School Certificate Science, etc) did not always produce the highest correlations with the criterion measure. For example, the Mathematics Test predicted School Certificate Social Studies better (0.86) than the Social Studies reference test (0.80). In fact, the Mathematics Test (excluding the combined Total Tests) proved to be the single most powerful predictor of class means, except for School Certificate English.

In contrast, the prediction of class standard deviations (Table 4.24) resulted in a series of low to moderate correlations. Just over a third of the correlations were not statistically significant, although this was more likely, due mainly to the Vocabulary Tests which were particularly short in length (only 12 items). A median correlation of 0.25 was recorded across the four Vocabulary Tests, 0.36 for the Essay Test, 0.49 for the Comprehension Tests, 0.54 for the combined Vocabulary/Comprehension Tests, 0.57 for the Mathematics Test and 0.67 for the GSAT. The

single most effective predictor of the class standard deviations was GSAT, although the Total Vocabulary/Comprehension Test (i.e. English, Science and Social Studies combined) correlated at a moderately high level with a median r of 0.73.

The rest of the results relating to the prediction of class performance - that is, the respective sub-analyses for the multiple-choice and open-ended/cloze item types - have been reported next in Tables 4.25 to 4.28.

TABLE 4.25

Prediction of School Certificate Class Means Using Multiple-Choice Tests. (N = 18)

Reference Test	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocabulary:					
English	(.88)	.85	.81	.67	.88
Science	.72	.72	(.75)	.58	.77
Social Studies	.87	.72	.82	(.60)	.85
Total	.92	.86	.87	.69	(.92)
Comprehension:					
English	(.71)	.81	.81	.60	.80
Science	.77	.81	(.82)	.66	.83
Social Studies	.81	.86	.82	(.68)	.87
Total	.82	.88	.86	.69	(.89)
Mathematics	.85	(.94)	.93	.80	.94

All r's are statistically significant at $p < 0.01$.

TABLE 4.26

Prediction of School Certificate Class Standard Deviations Using Multiple-Choice Tests. (N = 18)

Reference Test	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocabulary:					
English	(.48)	.41	.43	.29*	.60
Science	.53	.22*	(.24)*	.21*	.43
Social Studies	-.05*	-.05*	-.14*	(.05)*	-.01*
Total	.65	.44	.46	.35*	(.68)
Comprehension:					
English	(.63)	.66	.50	.41	.66
Science	.46	.61	(.45)	.53	.56
Social Studies	.57	.58	.41	(.67)	.54
Total	.72	.79	.58	.72	(.76)
Mathematics	.58	(.75)	.48	.46	.71

*Are not statistically significant; all other r's significant at $p < 0.05$

TABLE 4.27

Prediction of School Certificate Class Means Using Open-Ended/ Cloze Tests. (N = 17)+

Reference Tests	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocabulary:					
English	(.76)	.68	.76	.78	.79
Science	.68	.71	(.74)	.78	.76
Social Studies	.64	.75	.86	(.77)	.79
Total	.77	.76	.86	.81	(.82)
Comprehension:					
English	(.83)	.76	.83	.70	.86
Science	.73	.73	(.71)	.65	.78
Social Studies	.87	.83	.83	(.73)	.90
Total	.90	.83	.86	.74	(.92)
Mathematics	.72	(.95)	.86	.86	.91

All r's statistically significant at $p < 0.01$.

+An error in the administration of the tests resulted in one class being given the multiple-choice format only.

TABLE 4.28

Prediction of School Certificate Class Standard Deviations Using Open-Ended/Cloze Tests. (N = 17)

Reference Test	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocabulary:					
English	(.13)	.00	.02	.41*	-.01
Science	.08	.10	(.23)	.43*	.15
Social Studies	.10	.12	.01	(.37)	.09
Total	.14	.20	.24	.55*	(.21)
Comprehension:					
English	(.09)	-.01	-.11	.15	.08
Science	.05	.16	(.22)	.34	.21
Social Studies	-.12	.04	-.05	(-.00)	-.02
Total	.19	.17	.34	.38	(.25)
Mathematics	.65*	(.36)	.33	.16	.47*

*Are statistically significant at $p < 0.05$, all other r's are not significant.

A comparison of the correlations presented in Tables 4.23 and 4.25, showed no significant difference in the prediction of class means whether using a multiple-choice or open-ended/cloze

format. The median correlations obtained using the multiple-choice tests were 0.82 for Vocabulary and Comprehension, 0.88 for Vocabulary/Comprehension and 0.93 for Mathematics. With the corresponding open-ended/cloze tests the median correlations were 0.77 for Vocabulary, 0.81 for Comprehension, 0.87 for Vocabulary/Comprehension and 0.86 for Mathematics. The two sets of results also compare very closely to the combined analysis across item types, reported in Table 4.23.

Unlike the mean analysis, the use of different item types was shown to have a major effect on the prediction of class standard deviations. A comparison of Tables 4.26 and 4.28, clearly illustrates the discrepancy in the correlations obtained for each item type. With the multiple-choice tests, the median correlation for Vocabulary was 0.38, for Comprehension 0.61, for Vocabulary/Comprehension 0.68, and for Mathematics 0.58. In comparison, the median correlations for the open-ended/cloze tests were not even statistically significant, 0.12 for Vocabulary, 0.18 for Comprehension, 0.27 for Vocabulary/Comprehension and 0.36 for Mathematics. Interestingly, the multiple-choice sub-analysis shows this format to have produced higher correlations than the overall analysis (across item types), reported in Table 4.24. Possible reasons for this difference between the two item formats will be discussed in the next chapter.

The final set of results in this section focusses on the cloze Comprehension Test employing synonym replacement marking,

and how it compares in the prediction of class parameters using the exact replacement schedule, reported earlier in Tables 4.27 and 4.28. The relevant figures have been presented below in Tables 4.29 and 4.30.

TABLE 4.29

Prediction of School Certificate Class Means Using Cloze Tests With Synonym Replacement. (N = 17)

Cloze Reference Test	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Comprehension:					
English	(.92)	.81	.81	.73	.89
Science	.76	.75	(.71)	.68	.80
Social Studies	.87	.87	.87	(.76)	.92
Total	.92	.87	.88	.77	(.93)

All r's statistically significant at $p < 0.01$

TABLE 4.30

Prediction of School Certificate Class Standard Deviations Using Cloze Tests With Synonym Replacement (N = 17)

Cloze Reference Test	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Comprehension:					
English	(.16)	.18	-.10	.16	.18
Science	.00	.37	(.11)	.18	.25
Social Studies	-.11	.19	-.16	(-.04)	.02
Total	.32	.38	.26	.32	(.33)

All r's statistically not significant

A comparison of the mean prediction results reported in Tables 4.27 and 4.29 show a small, but consistent improvement in the correlations obtained with the cloze tests using synonym replacement marking. Although there was only a small difference in the respective medians of the overall Comprehension Test - 0.81 for cloze using exact replacement and 0.83 using synonym replacement - all but three of the correlations increased on average 0.034, with the largest gain occurring for the English Comprehension and School Certificate English correlation from 0.83 to 0.92.

Comparison of the standard deviation results in Tables 4.28 and 4.30 revealed a similar trend of small, but consistent increases in the correlations from the cloze employing synonym replacement marking. However, neither of the median correlations were statistically significant, 0.18 for cloze using exact replacement and 0.22 with synonym replacement; only one of the correlations was large enough to register as significant.

Multiple Regression Predictions of School Certificate Class Parameters

Multiple regression refers to the prediction of a criterion - in this instance, School Certificate Examinations - from two or more optimally combined independent variables - namely, the battery of reference tests. More specifically, the analysis again focusses on the prediction of mean and standard deviation parameters for each class.

The format of the multiple regression analysis, consisted firstly of a reanalysis of the Vocabulary/Comprehension Tests for each subject (in comparison to the simple correlations reported earlier); secondly, each of the Vocabulary/Comprehension Tests formed the basic equation for further analyses incorporating the Mathematics and GSAT Tests respectively, as additional predictor tests; thirdly, Total score predictions, combining tests across subjects, were also analyzed; and lastly, the Essay Test was incorporated into all analyses concerned with the prediction of School Certificate English, namely the English Vocabulary/Comprehension Test.

To assist with the presentation of the multiple-R's, the predictor tests have been reported in a short-hand notation as set out below:

Predictor Test	Notation	Predictor Test	Notation
English	E	Essay	Ey
Mathematics	M	GSAT	G
Science	S	Vocabulary	V
Social Studies	SS	Comprehension	C

These notations can stand alone or be used in combination, for example: English Comprehension = EC, or Science Vocabulary = SV.

The predictor tests have been listed in order of entry to the regression equation (employing a stepwise/forced entry regression programme) for the prediction of the corresponding School Certificate subject as indicated by the brackets.

The presentation of results follows the normal format of a combined analysis across item format, and then separate sub-analyses by item types. The mean and standard deviation results are also presented in individual tables. The first of these results - the combined analyses - have been reported next in Tables 4.31 and 4.32.

TABLE 4.31

Multiple-R Predictions of School Certificate Class Means Across Item Types. (N = 18)

Predictor Tests	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
EV; Ey; EC	(.90)	.87	.90	.82	.91
SC; SV	.83	.89	(.89)	.81	.90
SSV; SSC	.85	.91	.94	(.80)	.93
SSC; SSV; Ey; SV; EV; EC; SC	.92	.92	.97	.84	(.93)
Maths + GSAT: M; G	.86	(.98)	.93	.87	.96
Vocab + Comp + Maths:					
EV; Ey; M; EC	(.92)	.98	.95	.89	.98
M; SC; SV	.84	.99	(.94)	.87	.96
M; SSV; SSC	.85	.99	.96	(.87)	.97
M; EV; Ey; SV; EC; SC; SSV; SSC	.93	.99	.98	.92	(.98)
Vocab + Comp + GSAT:					
EV; Ey; EC; G	(.92)	.91	.90	.85	.93
SC; SV; G	.87	.92	(.91)	.83	.92
G; SSV; SSC	.87	.92	.95	(.82)	.94
SSC; SSV; Ey; SV; G; EV; EC; SC	.93	.95	.96	.86	(.95)

All R's statistically significant at $p < 0.01$

TABLE 4.32

Multiple-R Predictions of School Certificate Class Standard Deviations Across Item Types. (N = 18)

Predictor Tests	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
Ey; EC; EV	(.65)	.53	.60	.75	.69
SC; SV	.39	.57	(.44)	.61	.54
SSC; SSV	.31*	.49	.23*	(.58)	.37*
SC; Ey; SSV; EC; SV; EV; SSC	.73	.69	.71	.85	(.75)
Maths + GSAT:					
G; M	.72	(.72)	.61	.69	.74
Vocab + Comp + Maths:					
M; EV; EC; Ey	(.83)	.67	.64	.75	.80
M; SV; SC	.73	.68	(.53)	.61	.71
M; SSV; SSc	.73	.69	.45	(.62)	.69
M; EV; EC; Ey; SSV; SV; SSC; SC	.92	.73	.71	.85	(.82)
Vocab + Comp + GSAT:					
G; EC; Ey; EV	(.72)	.80	.69	.80	.81
G; SV; SC	.60	.73	(.62)	.73	.72
G; SSV; SSC	.59	.75	.63	(.75)	.71
G; EC; SSV; Ey; SV; SC; EV; SSC	.79	.83	.75	.88	(.83)

*These R's are statistically not significant; all others are significant at $p < 0.05$.

The multiple-R analysis predicting class means (Table 4.31) produced consistently high multiple correlations ranging from 0.80 to 0.99. All median correlations were 0.91 or higher, with Vocabulary/Comprehension recorded at 0.91, Mathematics/GSAT at 0.93, Vocabulary/Comprehension/Mathematics at 0.96 and Vocabulary/Comprehension/GSAT at 0.92. The standard deviation analysis (Table 4.32) produced multiple correlations mainly in the medium to medium high region, with coefficients ranging from 0.23 to 0.92. The median correlation for the Vocabulary/Comprehension Tests was 0.57, 0.72 for Mathematics/GSAT, 0.74 for Vocabulary/Comprehension/Mathematics and 0.77 for Vocabulary/Comprehension/GSAT.

The rest of the multiple-R results, namely the sub-analyses by item format have been reported below in Tables 4.33 to 4.36.

TABLE 4.33

Multiple-R Predictions of School Certificate Class Means Using Multiple-Choice Tests (N = 18)

Predictor Tests	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
EV; Ey; EC	(.90)	.94	.89	.72	.94
SC; SV	.79	.82	(.84)	.67	.85
SSC; SSV	.89	.86	.87	(.69)	.91
EV; EC; Ey; SV; SSV; SC; SSC	.93	.95	.91	.73	(.95)
Vocab + Comp + Maths:					
EV; Ey; EC; M	(.92)	.96	.93	.81	.96
M; SV; SC	.86	.94	(.94)	.78	.95
M; SSV; SSC	.91	.95	.94	(.78)	.96
M; EV; Ey; SV; SSV; SC; EC; SSC	.95	.97	.95	.83	(.97)

All R's statistically significant at $p < 0.01$.

TABLE 4.34

Multiple-R Predictions of School Certificate Class Standard Deviations Using Multiple-Choice Tests. (N = 18)

Predictor Tests	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
EC; Ey; EV	(.69)	.66	.71	.52	.81
SC; SV	.63	.61	(.47)	.53	.64
SSC; SSV	.58	.59	.41	(.72)	.57
EC;SV; SSV; SC; EV; Ey; SSC	.81	.81	.85	.92	(.90)
Vocab + Comp + Maths:					
EC; Ey; EV; M	(.73)	.80	.73	.57	.90
M; SC; SV	.68	.83	(.56)	.60	.78
SSC; SSV; M	.65	.78	.54	(.72)	.74
M; EV; Ey; SSC; SSV; SV; SC; EC	.81	.88	.88	.92	(.94)

All R's statistically significant at $p < 0.05$

TABLE 4.35

Multiple-R Predictions of School Certificate Class Means Using Open-Ended/Cloze Tests. (N = 17)

Predictor Tests	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
EC; Ey; EV	(.92)	.78	.88	.79	.90
SV; SC	.77	.78	(.79)	.79	.83
SSV; SSC	.87	.86	.91	(.81)	.92
SSC; Ey; SV; EV; EC; SSV; SC	.99	.89	.96	.84	(.96)
Vocab + Comp + Maths:					
EC; Ey; M; EV	(.93)	.96	.94	.91	.98
M; SC; SV	.79	.95	(.88)	.89	.93
M; SSV; SSC	.87	.95	.93	(.88)	.95
M; SSC; Ey; SV; EC; EV; SSV; SC	.99	.97	.97	.94	(.98)

All R's statistically significant at $p < 0.01$

TABLE 4.36

Multiple-R Predictions of School Certificate Class Standard Deviations Using Open-Ended/Cloze Tests. (N = 17)

Predictor Tests	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
Ey; EC; EV	(.40)	.20*	.50	.65	.44
SC; SV	.08*	.16*	(.26)*	.46	.22*
SSC; SSV	.14*	.14*	.05*	(.38)*	.09*
Ey; SSV; EC; SV; SSC; SC; EV	.60	.33*	.54	.74	(.47)
Vocab + Comp + Maths:					
M; EV; Ey; EC	(.78)	.40	.53	.66	.62
M; SC; SV	.66	.40	(.42)	.49	.52
M; SSV; SSC	.70	.45	.34*	(.47)	.53
M; SC; SSV; Ey; SV; EC; SSC; EV	.92	.54	.58	.77	(.68)

*Are statistically not significant, all other R's significant at $p < 0.05$.

A comparison of the multiple-R predictions for class means in Tables 4.31 and 4.33, revealed no significant difference in the respective sets of correlations when using multiple-choice or open-ended/cloze formats. A large proportion of the correlations in

both sets were high, in the 0.80's to 0.90's, with median correlations of 0.88 (Vocabulary/Comprehension) and 0.94 (Vocabulary/Comprehension/Mathematics) for the multiple-choice tests, and corresponding figures of 0.88 and 0.93 for the open-ended/cloze tests. However, the results of the standard deviation analysis revealed a different trend, consistent with earlier findings. The size of the multiple correlations reported in Tables 4.34 and 4.36 were clearly different. The multiple-choice format produced coefficients ranging from 0.41 to 0.94, with median correlations of 0.68 for the Vocabulary/Comprehension and 0.75 for Vocabulary/Comprehension/Mathematics. In contrast, the open-ended /cloze format resulted in some very low (i.e. not significant) and medium range coefficients, with respective median correlation figures of 0.34 and 0.57.

The last of the multiple-R analyses focused on the cloze Comprehension Test utilizing synonym replacement marking, and specifically, how it compares with the performance of the standard exact replacement schedule, as reported in Tables 4.35 and 4.36. These results have been presented below in Tables 4.37 and 4.38.

TABLE 4.37

Multiple-R Predictions of School Certificate Class Means Using Cloze Tests With Synonym Replacement. (N = 17)

Predictor Tests	School Certificate Means				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
EC; Ey; EV	(.97)	.83	.87	.82	.93
SV; SC	.79	.80	(.79)	.80	.85
SSV; SSC	.88	.90	.95	(.84)	.95
SSC; SV; SC; Ey; EV; EC; SSV	.98	.93	.98	.92	(.98)
Vocab + Comp + Maths:					
EC; Ey; EV; M	(.97)	.97	.94	.91	.98
M; SC; SV	.81	.95	(.88)	.89	.94
M; SSV; SSC	.88	.96	.95	(.88)	.96
SSC; SV; SC; Ey; EV; M; SSV; EC	.98	.98	.98	.94	(.98)

All R's statistically significant at $p < 0.01$.

TABLE 4.38

Multiple-R Predictions of School Certificate Class Standard Deviations Using Cloze Tests With Synonym Replacement. (N = 17)

Predictor Tests	School Certificate Standard Deviations				
	English	Mathematics	Science	Social Studies	Total
Vocab + Comp:					
Ey; EC; EV	(.43)	.28*	.51	.67	.46
SC; SV	.08*	.38*	(.24)*	.45	.27*
SSC; SSV	.15*	.21*	.16*	(.37)*	.09*
SV; EC; SSV; Ey; EV; SSC; SC	.53	.61	.76	.74	(.64)
Vocab + Comp + Maths:					
M; EV; EC; Ey	(.75)	.45	.54	.67	.61
M; SV; SC	.67	.50	(.41)	.48	.53
M; SSV; SSC	.69	.49	.35*	(.44)	.53
M; EV; SC; SV; EC; Ey; SSV; SSC	.79	.71	.78	.74	(.75)

*Are statistically not significant, all other R's significant at $p < 0.05$.

A comparison of the class mean analysis presented in Tables 4.35 and 4.37 show there generally to be only minor, insignificant fluctuations between the two sets of multiple correlations, except for the Vocabulary/Comprehension English Test, where more notable gains occurred especially with School Certificate English and Mathematics. The median results were very close, with figures of 0.87 (Vocabulary/Comprehension) and 0.93 (Vocabulary/Comprehension/Mathematics) for exact replacement marking, compared to corresponding medians of 0.89 and 0.95 for synonym replacement marking. A look at the standard deviation analyses in Tables 4.36 and 4.38, shows exactly the same pattern of results as described for the mean analysis above, an improvement for the first set of English reference test correlations, otherwise little or no change. Median multiple correlations of 0.26 (Vocabulary/Comprehension) and 0.57 (Vocabulary/Comprehension/Mathematics) were recorded for the exact replacement schedule, with corresponding figures of 0.38 and 0.59 for synonym replacement.

CORRECTION FOR SHRINKAGE IN THE MULTIPLE REGRESSION EQUATIONS

When multiple regression coefficients are calculated, the method of an optimal weighting of the predictor variables (i.e. the reference tests) tends to benefit from any chance effects available, causing the correlation estimates to be higher than the corresponding parameter R's. A formula developed by Wherry (cited in Carter, 1979) may be used to estimate the amount of shrinkage

expected in R^2 if the set of regression weights derived for the original sample were applied to a new sample. The formula used was:

$$\hat{R}^2 = 1 - (1-R^2)\left(\frac{n-1}{n-m-1}\right)$$

where $\hat{R}^2$ is an estimate of the parameter R;

R^2 is the square of the multiple correlation;

n is the number of cases; and

m is the number of predictor variables

Table 4.39 contains the results of this analysis with uncorrected and corrected multiple-R's for both multiple-choice and open-ended/Cloze formats.

TABLE 4.39

Multiple-R Predictions of School Certificate Class Means by Item Type (from Tables 4.31 and 4.33), Corrected for Shrinkage.

Reference Tests	Multiple-Choice		Open-Ended/Cloze	
	Uncorrected R	Corrected R	Uncorrected R	Corrected R
Vocab + Comp:				
English + Essay	.90	.88	.92	.90
Science	.84	.81	.79	.74
Social Studies	.69	.64	.81	.77
Total	.95	.91	.96	.93
Vocab + Comp + Maths:				
English + Essay	.92	.89	.93	.91
Science	.94	.92	.88	.85
Social Studies	.78	.73	.88	.85
Total	.97	.94	.98	.96
Median	.91	.89	.90	.88

As the difference between the median scores in Table 4.39 illustrates, the degree of shrinkage expected on a new sample was only .02 for both multiple-choice and open-ended/cloze formats. Thus, it can be said with reasonable certainty that only a very small part of the multiple-R relationships were due to error fitting (i.e. chance factors).

CHAPTER V

DISCUSSION OF RESULTS

The chief purpose of this chapter is to discuss the main findings and relate these to earlier studies of moderation, further technical consideration, and to offer explanations where appropriate. The discussion, of course, focusses mainly on the prediction of School Certificate class parameters. However, before turning to the issue of validity, brief comments on other aspects of the results - namely, parallelism of equivalent test forms, the experimental sample, sex differences and reliability - are presented first.

Parallelism of Equivalent Forms of the Reference Tests

Since it was intended from the outset to convert the reference test scores to standard scores to enable a direct comparison between the two item types (i.e. multiple-choice vs open-ended/cloze), only a minimal amount of time was allotted to the construction of the multiple forms of each test as strictly parallel, as in a traditional test development exercise (e.g. equating item difficulty). Rather, multiple test forms were incorporated as part of the multiple matrix sampling procedure, to enable the sampling of a number of different skill areas using different groups of pupils.

It is also, to a certain degree, more ecologically valid to demonstrate that the reference tests still discriminate sufficiently

well for moderation purposes when constructed under restricted conditions (as may be the case in practice). Thus, had the reference tests been constructed under optimal conditions, their performance in predicting the School Certificate Examinations would likely have shown further improvement.

Despite their less than optimal construction, the two sets of parallel forms were generally similar as indicated by the means and standard deviations. There was some minor variation, this being more noticeable in the open-ended/cloze test format.

Experimental Sample

The sample of fifth formers selected for this study was compared to the national population using percentage pass rates. There was little or no difference between the two sets of results, indicating that the performance of the experimental sample on the School Certificate Examinations was very similar to that of the national population of fifth formers (see Table 4.17).

That the sample does reflect the national population is particularly important for the external validity of the study. That is, the extent to which the findings of the study can be applied to the real world situation. Specifically, it gives greater validity to comparisons with earlier studies (see later in this chapter) and the partial replication of Gilmore's (1979) study.

Sex Differences

The investigation into sex differences (see Table 4.19) revealed several well-identified trends in relation to the School Certificate Examinations; namely, females performed better in English, while males did better in Mathematics, Science, Geography and History. Interestingly, no significant difference was detected for the School Certificate Total score (English plus best three). This pattern of results coincided exactly with that obtained nearly a decade earlier by Gilmore (1979, p 96).

Whatever the reasons for these well entrenched sex differences in the examinations (e.g. the different patterns of retention in the senior school for each sex), it was of interest to see if the same pattern was evident with the reference tests. Again, males did better on the Mathematics and Science tests, as well as GSAT; females performed better on the combined English/Essay Test; while there was no significant difference on the Social Studies Test.

All of the sex difference in the English/Essay Test was found to be confined to the Essay Test alone. Since the School Certificate English Examination consists largely of composition and comprehension of various kinds of prose, then the sex differences may be more attributable to the composition component than other parts. However, there is contrasting evidence from the PAT Reading Comprehension test where girls do better than boys.

Is it desirable to have reference tests with sex differences that correspond to similar differences identified in external examinations? The same criticism was recently levelled at ASAT, initiating a major study by ACER (Adams, 1984) to investigate possible causes of the difference in mean scores on ASAT between the two sexes. Adams' findings indicated that about half of the variance could be explained by the different retention rates at the senior school level; while the remaining variance was related to English ability, experience in mathematics and confidence in success. His final conclusion was that, "A student's sex had no significant direct effect on ASAT scores." (Ibid, p 106)

In relation to Adams' findings, the sex differences noted in the reference tests in this study are also probably due to the same sort of causes, rather than an inherent quality of the tests advantaging one sex over the other.

Reliability

The reliability estimates were calculated using the odd-even split-half method with a correction for length by the Spearman-Brown formula. The use of an internal consistency measure was necessitated when the MMS procedure left no opportunity for an alternate forms analysis, while time constraints similarly ruled out the possibility of a test-retest estimation. Such constraints also limited the present analysis to Form A of each test only. The split-half method provides a measure of consistency or equivalence with regard to content sampling, but no measure of stability over time.

To provide a clearer picture for discussion, all reliability estimates for the reference tests reported earlier have been summarized in Table 5.1 below.

TABLE 5.1

Summary of the Split-Half Reliability Coefficients of the Reference Tests for Combined, Multiple-Choice and Open-Ended/Cloze Formats.

	No of Items MC	(OE/C)	Combined (MC+OE/C)	Multiple -Choice	Open-Ended /Cloze	Synonym Replacement
Vocabulary:						
English	12	(12)	.70	.56	.80	-
Science	12	(12)	.62	.66	.65	-
Social Studies	12	(12)	.67	.60	.74	-
Comprehension:						
English	17	(49)	-	.66	.89	.94
Science	17	(44)	-	.74	.91	.93
Social Studies	17	(50)	-	.72	.79	.91
Vocab + Compre:						
English	29	(61)	-	.76	.97	.99
Science	29	(56)	-	.74	.94	.96
Social Studies	29	(62)	-	.81	.84	.95
Mathematics	25	(25)	.88	.86	.88	-
GSAT	25	-	-	.78	-	-
Essay*	-	(1)	-	-	.64	-

*Inter-rater coefficient only

The reliability coefficients, ranging from a low of 0.56 to a high of 0.99, must generally be considered as most satisfactory in relation to the modest length of the respective tests. For purposes of group prediction, as against individual assessment, Kelley's (1927) recommended minimum reliability criterion of 0.50 for making placement decisions about groups of pupils, still retains its importance today (e.g. Sax, 1980). Using Kelley's figure as a guide, it is clearly evident from the reliability estimates shown in Table 5.1 that the reference tests achieved this minimum standard. In fact, the majority of the coefficients were much higher than the 0.50 figure, and those which are serious candidates for actual use in a national moderation system (i.e. Vocabulary/Comprehension, Mathematics and GSAT) were all above 0.76.

A comparison of the reliability estimates for the multiple-choice and open-ended/cloze formats demonstrated consistent superiority for the latter in the Vocabulary, Comprehension and Vocabulary/ Comprehension Tests, but not for the Mathematics Tests. The probable reason for this difference was the increased number of items in the cloze format of the Comprehension Test (44 to 50 and 56 to 62 items) compared with the much shorter multiple-choice format (17 and 29 items). Consistent with this trend was the Mathematics Test, where both test formats had the same number of items resulting in almost identical reliability coefficients (.86 vs .88). Thus, if the multiple-choice Comprehension Tests had been of similar length as the cloze format, and other things being equal (e.g. item quality), then it also would have produced similarly high estimates of reliability.

Against this trend, however, were the Vocabulary coefficients where the open-ended format had higher overall estimates, yet both formats had an equal number of items. There was evidence here to suggest that this was an example of "other things not being equal", namely, that Form B of the multiple-choice English Test had a substantially higher validity coefficient than Form A (r's = .57 vs .72). In addition, the Form A estimate of 0.56 appears somewhat lower in comparison to the other open-ended figures, as does the Social Studies estimate of 0.60. Possibly, these two test forms have been affected by technical deficiencies in the quality of the items, not detected during the pilot testing.

The coefficients for the Vocabulary/Comprehension Test (Table 5.1) again illustrate the principle of "the longer the test, the higher the reliability". When the two relatively short tests were combined the reliability estimates were improved. This was consistent for both test formats. In the context of moderation, it is important for the reference tests to be relatively short, while still possessing adequate reliability. It may be helpful therefore, to optimize this relationship by using the Spearman-Brown formula in reverse to specify an increase in the length of a test, and then to estimate the appropriate level of reliability (Guilford and Fruchter, 1978, pp 425-6). For example, by taking the short Vocabulary Tests and applying them to the Spearman-Brown formula thus:

$$r = \frac{r \times y}{\sqrt{\frac{1 - r}{n} + r}}$$

and making the assumption of doubling the length of the test (2n); then the correction for test length would have resulted in reliability estimates of 0.78 for English Vocabulary (0.70 uncorrected), 0.79 for Science Vocabulary (0.62) and 0.75 for Social Studies (0.67).

The use of synonym replacement marking with the cloze format of the Comprehension Test generally improved the reliability estimates (see Table 5.1). The gains ranged from .02 to .05 for the English and Science Tests, while larger increases of .11 and .12 were recorded for the Social Studies Tests. The large improvement for the latter tests was probably a reflection of the greater difficulty level of these tests (see Table 4.10), resulting in an increase in the proportion of pupils supplying suitable alternative answers, as against correct answers.

The essay inter-rater reliability coefficient of 0.64 was only a modest outcome. However, in view of the evidence about the unreliability of essay marking, it was probably not an unduly low or unexpected figure for two untrained, novice markers of fifth form English essays. Elley, Barham, Lamb and Wyllie (1979, pp 88-95) reported a median r of 0.68 for a group of 16 experienced English teachers who "were briefed carefully on the criteria and marking patterns to be adopted". (Ibid, p 88). They also practised on "guinea-pig" scripts. Thus, the median r of 0.64 by two novice markers in the current study, would not have been out of place amongst the efforts of experienced English teachers from Elley,et al's study (1979).

PREDICTIVE VALIDITY

The correlations between the reference tests and the School Certificate Examinations represent a form of predictive validity. That is, an indication of the ability of the reference tests to predict the School Certificate Examinations. Thus, by demonstrating that the reference tests correlate highly with the School Certificate results, then it can be argued that the reference tests are capable of acting as a moderating device for Sixth Form Certificate subjects in the same way the School Certificate Examinations are used currently; but without the many disadvantages associated with external examinations.

It is also intended here, to compare the results with those reported in earlier, local studies of moderation.

Again, it should be emphasized that the combined Vocabulary/ Comprehension Test is the important format under evaluation as a moderating instrument. Therefore, the presentation of separate results for the Vocabulary and Comprehension Tests should be considered as of academic interest only.

Correlations Between the Reference Tests and School Certificate Marks

This analysis was based on pupils' individual scores and its main function was to enable a comparison to be made with two earlier studies, for which class-based predictions were not conducted.

The overall results, across item types, (Table 4.20) produced a series of moderately high correlations - over 80 percent of the r's were 0.60 or higher - that were also quite stable when correlated with different subject tests. Generally, the corresponding subject reference test correlated the highest with the appropriate School Certificate Examination, while the combining of the relatively short Vocabulary and Comprehension Tests resulted in moderately high correlations (median r = .73). The Mathematics Test also predicted well (median r = .72). A comparison of the separate sub-analyses by item type (Tables 4.21 and 4.22), revealed little or no difference in the correlations whether using the multiple-choice or open-ended/Cloze format. The final point of interest was that the multiple-choice Comprehension format correlated slightly better with the cloze using synonym replacement marking, than with exact replacement only. This would appear to contradict previous research (e.g. Bormuth, 1965a, 1965b; Hargis, 1972; McKenna, 1976; and Elley, 1976) where it has generally been reported that little or no significant difference exists between the two scoring procedures. However, Henk (1981), has also found a greater degree of variation between the two scoring procedures in line with the current study.

Table 5.2 provides a comparison of results from some earlier studies. In the case of the current study and Gilmore's, only the correlations between the corresponding reference test and appropriate School Certificate Examination have been presented. Hulbert's Experimental Moderating Test (EMT) format was a multiple-choice

comprehension test using passages from several subject areas. He did not develop specific subject tests. Elley and Livingstone's figure was based on PAT Reading Comprehension scores collected on entry to secondary school, thus predicting from a point over two years earlier. In the other studies all data collection was conducted during the fifth form year.

TABLE 5.2

A Comparison of Results from Related Studies Showing Correlations Between Reference Tests and School Certificate Examination Marks.

S.C. Subjects	Chamberlain (1988)	Gilmore (1979)	Hulbert (1978)	Elley and Livingstone (1972)
	SRT*	SRT*	EMT	PAT Rding. Comp.
English	.69+	.66	.66	.75
Mathematics	.79	.78	-	-
Science	.76+	.62	.58	-
Social Studies	.74+	-	.55	-
Median	.75	.66	.58	.75

*SRT = Subject-based Reference Tests

+Combined Vocabulary/Comprehension Test

The correlations from the current study compared very favourably with those from previous studies. Although numerous

factors may account for the difference in the median correlation scores, the Vocabulary/Comprehension format appeared to correlate with School Certificate results as well, if not better, in contrast with the other test contents tried. The case for the use of comprehension was further strengthened by the PAT Reading Comprehension result, showing a relatively high correlation (0.75) with School Certificate English two years later, and Hulbert's correlation of 0.66 also with School Certificate English. However, the use of specific subject tests in the current study and that of Gilmore, produced higher correlations than the general EMT format of Hulbert's.

The results in Table 5.2 were based on tests of developed-abilities, but in an additional analysis, the same four studies also investigated the use of scholastic aptitude tests as moderating instruments. Gilmore developed a multiple-form aptitude test specially for her study; while Hulbert and Elley and Livingstone reported on commercially published tests used by the schools. The current study utilized one form of Gilmore's test, thus acting as a replication study for this test. The resulting correlations with the School Certificate Examinations are presented below in Table 5.3.

TABLE 5.3

A Comparison of Results from Related Studies Showing Correlations Between Scholastic Aptitude Tests and School Certificate Examination Marks.

S.C. Subjects	Chamberlain (1988)	Gilmore (1979)	Hulbert (1978)		Elley and Livingstone (1972)	
	GSAT	GAT	OTIS	DAT-VR*	OTIS	DAT-VR*
English	.59	.72	.76	.72	.81	.71
Mathematics	.61	.65	.60	.52	.54	.44
Science	.65	.65	.61	.62	.47	.44
Social Studies	.64	.59	.53	.52	.55	.52
Median	.63	.65	.61	.62	.55	.48

*DAT - Verbal Reasoning sub-test.

The correlations between the various scholastic aptitude tests and the School Certificate Examination marks resulted in median r's ranging from 0.48 to 0.65. In comparison, the developed abilities tests produced median r's ranging from 0.57 to 0.75. The aptitude tests generally correlated higher with School Certificate English, but lower with Mathematics compared with the developed abilities tests. The use of specific-subject tests in the current study, and to a lesser extent by Gilmore, seem to produce the better overall result.

Correction for Attenuation

While the correlations between the reference tests and the School Certificate Examination marks were positive and moderately high, it should be remembered that the reliability of the reference tests, especially for the multiple-choice format (see Table 5.1), was less than perfect. This unreliability (or error of measurement) has the effect of lowering the correlations between these tests and their criterion. However, it is possible to estimate the "true" correlation by assuming the reference tests to have perfect reliability. The procedure of correcting for attenuation (i.e. unreliability) is calculated by dividing the correlation by the square root of the reliability coefficient, thus:

$$r_c = \frac{r\ xy}{\sqrt{r}}$$

A selection of uncorrected correlations (from Tables 4.21 and 4.22) and their corresponding corrected estimates, for both multiple-choice and open-ended/cloze formats, have been reported below in Table 5.4.

TABLE 5.4

Correlations Between the School Certificate Examination Marks and Reference Tests, Corrected for Attenuation.

Reference Test	S.C. Examination	Multiple-Choice		Open-Ended/Cloze	
		Uncorrected r	Corrected Estimate	Uncorrected r	Corrected Estimate
English Vocab	English	.65	.87	.61	.69
English Comp	English	.57	.70	.65	.69
Science Vocab	Science	.55	.68	.61	.75
Science Comp	Science	.59	.69	.56	.59
Social St Vocab	Social Studies*	.59	.76	.64	.74
Social St Comp	Social Studies*	.62	.73	.59	.66
Mathematics	Mathematics	.79	.85	.80	.85

*Combination of Geography, History and/or Economic Studies.

For the multiple-choice tests, the median correlation increased from 0.59 (uncorrected) to 0.73 (corrected estimate), with the largest gain occurring in English. In contrast, the open-ended/cloze tests produced a smaller gain in the median correlation from 0.61 to 0.69, with no significant subject variations. In essence, this result reflected the trends in the reliability coefficients reported earlier, (see Table 5.1), which demonstrated the higher reliability achieved with the open-ended/cloze format relative to a comparable multiple-choice format.

Prediction of School Certificate Class Parameters

The next two sections represent the main purpose of the current study. In the first section, the analysis of the results focusses on the correlation of the means and standard deviations of each class (N = 18) on the reference tests, with their corresponding means and standard deviations on the School Certificate Examinations. The expectation at the beginning of the study, based on Gilmore's (1979) findings, was that analysis at the group level would produce significantly higher correlations than those based on pupils' individual scores. The important assumption with group-based analysis) is the smaller probability for error to arise since individual pupils can change rank order within their class without greatly affecting the mean and standard deviation scores, and without any substantial change in the rank order of the classes.

As expected, the prediction of class means produced higher correlations (Tables 4.23, 4.25 and 4.27) than did the individual

pupil correlations with the School Certificate Examination marks. Median correlations of 0.83 (Vocabulary), 0.87 (Comprehension), 0.90 (Vocabulary/Comprehension), 0.91 (Mathematics) and 0.87 (GSAT) were recorded for the analysis across item types. The use of diferent test formats, either multiple-choice or open-ended/cloze, proved to have no significant effect on the size of the correlations reported for the combined analysis above. A comparison of these results with a similar analysis reported by Gilmore (1979) is presented below in Table 5.5. Again, only the correlations between the corresponding reference tests and appropriate School Certificate Examinations have been presented. In addition, correlations for the scholastic aptitude tests have also been provided.

TABLE 5.5

A Comparison of Two Studies Showing Correlations Between the Class Means on the Reference Tests and the School Certificate Examinations.

Subject	Chamberlain (1988)		Gilmore (1979)	
	SRT*	GSAT	SRT*	GAT
English	.87+	.86	.94	.94
Mathematics	.97	.90	.93	.89
Science	.89+	.87	.87	.91
Social Studies	.80+	.79	-	.87
Essay	.62	-	.69	-
Median	.87	.87	.90	.90

* Subject-based Reference Tests. All r's sig. at $p < 0.01$

\+ Combined Vocabulary/Comprehension Test.

Overall, the figures were very similar as can be seen from the respective median correlations of 0.87 and 0.90. The biggest differences occurred in the case of the English and Essay Tests, while the slightly higher correlations reported by Gilmore were actually lower in comparison to those of the current study when calculated at the individual level of analysis (see Table 5.2).

The prediction of the class standard deviation scores was notably less successful, resulting in a series of low to moderate correlations (see Table 4.24). A number of the correlations, limited mainly to the Vocabulary Tests, were not statistically significant. This was largely a reflection of the shortness of the Vocabulary and, to a lesser extent, the Comprehension Tests. GSAT produced the highest correlations and proved to be the best single predictor test. Just as for the class mean analysis, a comparison of the standard deviation results with those reported by Gilmore have been presented below in Table 5.6.

TABLE 5.6

A Comparison of Two Studies Showing Correlations Between the Class Standard Deviations on Reference Tests and the School Certificate Examinations.

Subject	Chamberlain (1988)		Gilmore (1979)	
	SRT*	GSAT	SRT*	GAT
English	.51+	.58	.48	.57
Mathematics	.57+	.70	.30**	.42
Science	.39	.61	.53	.54
Social Studies	.31+**	.67	-	.54
Median	.45	.64	.48	.54

* Subject-based Reference Tests

\+ Combined Vocabulary/Comprehension Test

All r's statistically significant at $p < 0.05$, except those marked by a **

The comparison of results clearly showed the general ability tests to be the better and more consistent predictor of class standard deviations than the subject-based reference tests. Median correlations of 0.54 and 0.64 were recorded for the general ability tests in comparison to median correlations of 0.45 and 0.48 for the subject-based reference tests. A comparison of the two sets of data showed the correlations generated by the respective subject-based reference tests to be very similar (median r's of 0.45 vs 0.48), while the GSAT correlations from the

current study (median r = 0.64) were generally higher than those reported by Gilmore (median r = 0.54).

Of significant interest was a major discrepancy in the ability of the two item types to predict the class standard deviation (see Tables 4.26 and 4.28). The multiple-choice format produced median correlations ranging from 0.38 to 0.68, while the open-ended cloze format resulted in only low, non-significant median r's. Further discussion of this discrepancy follows in a comparison of the item formats later in the chapter.

In line with the earlier results, the use of synonym replacement marking in the prediction of class parameters for the School Certificate Examinations, again resulted in small but consistent gains in the correlations (see Tables 4.29 and 4.30). Apparently, this scoring procedure provides a better measure of reading comprehension than the use of exact or verbatim scoring only. However, any technical improvements must be weighed against the additional labour, time, cost and unreliability associated with this type of scoring when thousands of candidates are involved. Further discussion on this issue also follows later in the chapter.

Multiple Regression Predictions

In the second and final section of the analysis for the prediction of class parameters, a series of multiple regression equations was generated to determine the best possible correlation for two or more of the reference tests in an optimally combined weighting. The following test combinations were analyzed:

1) Vocabulary/Comprehension; 2) Mathematics/GSAT; 3) Vocabulary/Comprehension/Mathematics; and 4) Vocabulary/Comprehension/GSAT. In addition, the Essay was added to all of the English Test analyses.

The multiple-R predictions of the School Certificate class means (see Table 4.31) produced consistently high correlations, with median R's ranging from 0.90 to 0.96. The separate sub-analyses by item type, revealed no significant difference in the size of the multiple-R's obtained with either the multiple-choice or open-ended/cloze test formats (see Tables 4.33 and 4.35). The median correlations were 0.88 and 0.94 for the former, and 0.87 and 0.93 for the latter.

The best subject-based prediction equations for the corresponding School Certificate Examination and the best optimal prediction equations (regardless of subject content) for the prediction of School Certificate Examination class means have been summarized in Table 5.7. These results are taken from the "across-item-type" analysis (see Table 4.31). As in the case of earlier sections of this chapter, appropriate multiple-R results from Gilmore's study have also been included as a comparison.

As can be seen from Table 5.7 the use of the appropriate subject tests only, as predictors of School Certificate Examination class means, resulted in R's ranging form 0.80 to 0.98, with a median of 0.90. Clearly, these correlations are sufficiently high in themselves to act as group discriminators for moderation purposes.

TABLE 5.7

Summary of the Multiple-R Predictions of the School Certificate Examination Class Means Across Item Types, With a Comparison of Gilmore's (1979) Data.

S.C. Subject	Subject Ref. Tests	R	R^2	SEE	Optimal Predictors	R	R^2	SEE	Gilmore's Optimal R
English	Engl. Vocab + Comp + Essay	.90	.81	3.67	Engl. Vocab + Comp + Essay + Maths	.92	.85	3.35	Engl + GAT = .96
					Engl. Vocab + Comp + Essay + GSAT	.92	.85	3.40	
Mathematics	Mathematics	.97	.94	2.05	Scie. Vocab + Comp + Maths	.99	.98	1.63	Maths + Scie= .94
					SoSt. Vocab + Comp + Maths	.99	.98	1.48	Maths + GAT = .94
					Maths + GSAT	.98	.96	1.78	
Science	Scie. Vocab + Comp	.89	.79	4.00	SoSt. Vocab + Comp + Maths	.96	.92	2.61	Maths + GAT = .92
					SoSt. Vocab + Comp + GSAT	.95	.90	2.96	
					Engl. Vocab + Comp + Essay + Maths	.95	.90	3.00	
					Scie. Vocab + Comp + Maths	.94	.88	3.24	
					SoSt. Vocab + Comp	.94	.88	2.98	
					Scie. Vocab + Comp + GSAT	.90	.81	3.82	
Social Studies	SoSt. Vocab + Comp	.80	.64	5.26	Engl. Vocab + Comp + Essay + Maths	.89	.79	4.38	GAT + Essay =
					SoSt. Vocab + Comp + Maths	.87	.76	4.59	GAT + Maths = $\bar{Mdn}$ R=.90
					Scie. Vocab + Comp + Maths	.87	.76	4.48	GAT + Engl =
					Maths + GSAT	.87	.76	4.43	
					SoSt. Vocab + Comp + GSAT	.82	.67	5.22	
Total	Vocab + Comp + Mathematics	.98	.96	8.20	Engl. Vocab + Comp + Essay + Maths	.98	.96	6.97	
					SoSt. Vocab + Comp + Maths	.97	.94	8.96	
					Scie. Vocab + Comp + Maths	.96	.92	9.69	N/A
					Maths + GSAT	.96	.92	9.60	
					Vocab + Comp + GSAT	.95	.90	13.78	

All R's statistically significant at $p < 0.01$

However, the addition of either the Mathematics or GSAT Test to the other subject-based tests produces an improvement in the multiple-R's, as the median correlation of 0.95 for the optimal predictors demonstrated. Although, some of the subject-based reference tests did not always produce the highest optimal prediction, in all cases their respective correlations were so close as to make other essential criteria - namely, face and content validity of the tests - more important in the selection of the appropriate predictor tests. For example, the best subject-based prediction for School Certificate Science was 0.94 (Science Vocabulary/Comprehension and Mathematics) which was only marginally less superior than the highest optimal prediction of 0.96 (Social Studies Vocabulary/Comprehension and Mathematics). No doubt, few science teachers would appreciate having their classes allocated grades on the basis of a Social Studies or English reference test, despite any reassurances about its technical attributes!

It could be argued that the addition of the Mathematics or GSAT Tests, in themselves may not be appreciated by teachers of other subjects. However, as the single most reliable and powerful of the short predictor tests overall, the use of either of these two tests in combination with the subject tests appears to produce a significant gain in the multiple-R that justifies their inclusion in time and resources. Certainly it would not be difficult to argue the case for the inclusion of a mathematics test to predict science, geography or economics courses, and similarly, a general

ability test to predict English and history courses of the senior secondary school level. In essence, the extra tests would be acting as a supplementary measure of reasoning in verbal contexts.

The group of lower multiple-R's (range .82 to .89) recorded for the prediction of the School Certificate Social Studies can be explained firstly, by the combining of the Geography, History and Economics Examinations into a single criterion measure. Had resources permitted individual analyses of these subjects, better results may well have been obtained. Gilmore, for example, analysed these subjects separately (although she did not have corresponding subject-based predictor tests), and produced multiple-R's of 0.91 for Geography, 0.95 for History and 0.88 for Economic Studies.

Secondly, the class size for each of these subjects was on average significantly smaller than in the more popular subjects. A reanalysis of the multiple-R's, dropping two classes with N's less than 10, resulted in the correlations increasing by an average of 0.023.

The final point of interest from Table 5.7 was the comparison with Gilmore's (1979) results. The respective figures were quite comparable overall, with median R's of 0.95 for the current study and 0.94 for Gilmore.

The multiple-R predictions of the School Certificate class standard deviations (see Table 4.32) produced moderate to moderately high correlations, with median R's ranging from 0.57

to 0.77. Consistent with the earlier trend, the multiple-choice format was found to correlate more highly than the open-ended/cloze format (see Tables 4.34 and 4.36). Median R's for the former were 0.68 and 0.75, while for the latter they were 0.34 and 0.57.

As for the analysis of means Table 5.8 contains a summary of the best subject-based and optimal prediction equations for the prediction of School Certificate Examination class standard deviations. These results are from the across-item-type analysis (see Table 4.32). The results from Gilmore's study have again been included.

Although the prediction of the class standard deviation is only of secondary importance in relation to predicting the class ability level as indicated by the mean score, it is still very useful to have an indication as to the range of abilities in each class. Clearly, a review of Table 5.8 reveals that the prediction of the standard deviation was a more difficult task. The corresponding subject-based reference tests produced correlations ranging from 0.44 to 0.82, with a median of only 0.57. However, the addition of the GSAT (the single most powerful predictor of the standard deviation) or Mathematics Tests for the optimal prediction analyses produced correlations ranging from 0.62 to 0.83 and a median R of 0.75.

Gilmore (1979) also found it difficult to predict the class standard deviation and resorted to using mean deviation scores

TABLE 5.8

Summary of the Multiple-R Predictions of the School Certificate Examination Class Standard Deviations Across Item Types, With a Comparison of Gilmore's (1979) Data.

S.C.Subject	Subject Ref. Tests	R	R^2	SEE	Optimal Predictors	R	R^2	SEE	Gilmore's Optimal-R
English	Engl Vocab + Comp + Essay	.65	.42	2.06	Engl Vocab + Comp + Essay + Maths Scie Vocab + Comp + Maths SoSt Vocab + Comp + Maths Engl Vocab + Comp + Essay + GSAT Maths + GSAT	.83 .73 .73 .72 .72	.69 .53 .53 .52 .52	1.58 1.87 1.85 1.97 1.82	Dev* + GAT + Engl = .73
Mathematics	Mathematics	.57	.33	2.48	Engl Vocab + Comp + Essay + GSAT SoSt Vocab + Comp + GSAT Scie Vocab + Comp + GSAT Maths + GSAT	.80 .75 .73 .72	.64 .56 .53 .52	1.99 2.15 2.21 2.17	Dev + GAT = .63
Science	Scie Vocab + Comp	.44	.19	3.12	Engl Vocab + Comp + Essay + GSAT Engl Vocab + Comp + Essay + Maths SoSt Vocab + Comp + GSAT Scie Vocab + Comp + GSAT	.69 .64 .63 .62	.48 .41 .40 .38	2.70 2.88 2.80 2.83	Dev + GAT + Scie = .80
Social St.	SoSt Vocab + Comp	.58	.34	3.58	Engl Vocab + Comp + Essay + GSAT Engl Vocab + Comp + Essay + Maths Engl Vocab + Comp + Essay SoSt Vocab + Comp + GSAT	.80 .75 .75 .75	.64 .56 .56 .56	2.80 3.10 3.01 2.97	Dev + Engl = Dev = Mdn R = .65 GAT =
Total	Vocab + Comp + Mathematics	.82	.67	9.02	Vocab + Comp + GSAT Engl Vocab + Comp + Essay + Maths Engl Vocab + Comp + Essay + Maths Vocab + Comp	.83 .81 .80 .75	.69 .66 .64 .56	8.78 7.58 7.89 9.94	N/A

All R's statistically significant at $p \leq 0.05$.

*Dev. = Mean deviation score

as predictors, in addition to her other reference tests. The mean deviation score proved to be quite powerful and was entered into most of the regression equations. However, as shown in Table 5.8, the multiple-R's obtained were - except for Science - lower than those produced in the current study. The median R for Gilmore was 0.69 compared with 0.75 in this study.

TOTAL SCORE PREDICTIONS

Throughout the Results and the current chapter, correlations have been reported for total scores on both the Vocabulary and Comprehension Tests (i.e. English, Science and Social Studies sub-tests combined), and for the School Certificate Examinations (i.e. English plus best three other subjects). Generally, these results have not been commented upon since the main focus of the study was the prediction of specific School Certificate subjects with corresponding subject reference tests. However, it was decided to include these additional results, partly because it provided a further alternative for consideration as a moderating scheme, and also because the computer enabled the extra data to be analysed without much additional workload for the researcher. Although the current policy (Ross Report, 1986) for the Sixth Form Certificate favoured moderation that is independent of other subjects, this approach still warrants further consideration in the light of the Australian experience with ASAT and Hulbert's (1978) research.

Hulbert's EMT derived a total score using a series of prose passages from various subject disciplines. Of greater interest is the ACER's ASAT test, which is currently employed to moderate group-one leaving certificate subjects at the senior secondary school level in two Australian states. ASAT also derives a single total score, even though the test contains a variety of material drawn from mathematics, science, social science and the humanities. Certainly from the evidence cited earlier (see Chapter II), this technique has proven to be technically sound and has worked well.

The correlations between the total score of the Vocabulary/ Comprehension Test, and the School Certificate Examinations were consistently .01 - .02 higher on average than for the best optimal correlation by any of the shorter subject tests. The multiple-R predictions of class means and standard deviations using total scores have been summarized below in Table 5.9.

TABLE 5.9

Summary of the Multiple-R Predictions of School Certificate Examination Class Means and Standard Deviations Using Vocabulary/ Comprehension Total Scores, Across Item Types. (N = 18)

Reference Tests	School Certificate				
	English	Mathe-matics	Science	Social Studies	Total
Class Means:					
Vocab + Compreh	.92	.92	.97	.84	.93
Vocab + Comp + Maths	.93	.99	.98	.92	.98
Vocab + Comp + GSAT	.93	.95	.96	.86	.95
Class Std Devs:					
Vocab + Compreh	.73	.69	.71	.85	.75
Vocab + Comp + Maths	.92	.73	.71	.85	.82
Vocab + Comp + GSAT	.79	.83	.75	.88	.83

All R's statistically significant at $p < 0.01$

A comparison of the multiple-R's reported in Tables 5.7 and 5.8 shows the total score correlations to be at least as high, if not marginally higher than those obtained with the individual subject tests. The gains were more significant for the prediction of class standard deviations with median differences of .06 (English), .07 (Science) and .10 (Social Studies).

The fact that a single test drawing material from several different subject areas can correlate so highly with each of the School Certificate Examinations serves to illustrate that there is a significant degree of inter-correlation among the academic subjects. Thus, despite the differences in subject content, clearly there was a common group of general skills and abilities being tested across subject divisions. This means that a large number of pupils who did well in one subject also tended to do well in the other subjects.

COMPARISON OF ITEM TYPES

Several interesting findings emerged from the comparisons of the two item formats under consideration here, namely the multiple-choice and open-ended/cloze item types.

Firstly, the open-ended/cloze format produced consistently higher reliability estimates (see Table 5.1). In the case of the cloze Comprehension Tests this was explained by the greater number of items; however, for the open-ended Vocabulary Tests - which had an equivalent number of items as the multiple-choice tests - a suitable explanation was not so obvious.

Secondly, the two item formats predicted School Certificate Examination class means equally well. There were slight fluctuations in the correlations, but generally, the variations were not significant or consistent. The only notable difference was that the cloze correlated slightly, but consistently better with School Certificate English. There is evidence (Ratnamalar, 1986)

that this difference is specific to differences in subject content. Ratnamalar found similar differences when using cloze tests to assess the readability of textbooks. She reported a higher mean score in English than for science or history. Thus, there appear to be important subject differences in the use of the cloze format that are worthy of further investigation.

Thirdly, the multiple-choice format predicted School Certificate Examination class standard deviations at a far superior level to the open-ended/cloze format. For the latter, the worst set of correlations occurred with the very short Vocabulary Test. However, the median correlations for the other tests, or combination of tests, were also not statistically significant. Why the open-ended/cloze format should not be able to discriminate standard deviation scores is difficult to determine. Possibly, it is simply the process of responding. That everyone selects an answer for multiple-choice tests, tends to create a group who answer correctly because they know already, a second group who make an "educated guess" using the answer options as cues to recall - perhaps half of whom may guess correctly, and a third group who simply do not know. This response strategy may be more sensitive in detecting the small standard deviation scores than the open-ended/cloze format where there was a much larger group who answered incorrectly since there were no cues provided to activate recall.

Whatever the reason, it was clear from the current study that given the same test format as used here, the open-ended and cloze item types were not sensitive enough, as used in this study,

to detect class standard deviation scores for moderation purposes. However, the multiple-R's generated appeared as reasonably promising and warrant further investigation.

There is obviously a need for flexibility in a moderating system in this area. It has been shown by Gilmore (1979) and the current study, that estimating the spread of ability (i.e. standard deviation) of each class is a significantly more difficult task than estimating the ability level (i.e. mean) of each class.

As a compromise, teachers assessments of their class distributions could be used as a guide, in addition to the reference test estimations of the standard deviation. Thus, a teacher would be in a position to "bargain" the composition (i.e. number of 1's, 2's 3's etc) of the grade allocation that has been made for his/her class. For every grade the teacher moved upwards there would need to be a corresponding grade that moved downwards. For example if a teacher felt that his/her class contained three pupils who clearly all deserved a Grade 1, but had been initially allocated only two Grade 1's, then he/she could only "bargain" for another Grade 1 by changing, say, one Grade 2 into a Grade 3, or two Grade 3's into two Grade 4's, depending on the distribution of grades that the particular school has been allocated.

Cloze Scoring: Exact vs Synonym Replacement

The use of synonym replacement in the scoring of the cloze Comprehension Tests resulted in small but consistent gains in the

correlations between the reference tests and the School Certificate Examinations. In many cases the increases were so minor as to be insignificant, but on other occasions - particularly where there was a suggestion that the passages were too difficult - the use of synonym replacement produced much larger and very significant gains in the correlations. This technique may be of some benefit in the classroom or research situation where a test is constructed with prose passages which prove to be too difficult.

Also of interest was that synonym replacement scoring resulted in higher reliability estimates than with exact replacement. This appeared to contradict some of the earlier findings (e.g. Hargis, 1972 and McKenna, 1976).

However, any slight gains in terms of the predictive validity of the reference tests must always be considered in relation to the increased time and resources required for synonym replacement scoring. The use of synonym replacement marking would make it impossible for the tests to be marked by a computer. However, when marked manually, it would probably only require about half as much time again to mark using synonym replacement in comparison to exact replacement only.

Multiple Regression Predictions vs Simple Correlation Predictions

For reasons of practical application, it may be of considerable importance to have a simpler method of calculating the correlations between the reference tests and the criterion measure, in this

case the School Certificate Examinations. If schools were to become responsible for their own moderation, the use of a multiple-regression programme may be a less viable alternative than one based on simple correlations of combined, unweighted variables.

Therefore, a comparison of the respective multiple-R's and simple r's in each subject has been presented below in Table 5.10.

TABLE 5.10

A Comparison of Multiple-R and Simple r Correlations.

S.C. Examination	Predictor Tests	R	r
English	EV + EC + Ey + M	0.92	0.92
Mathematics	M + G	0.98	0.97
Science	SV + SC + M	0.94	0.93
Social Studies	SSV + SSC + M	0.87	0.83

As can be seen from Table 5.10, there was generally little or no difference between the two sets of correlations. In all cases, where a difference was reported, the multiple-R produced the higher correlation. The English correlations were identical, while for mathematics and science there was a small difference of .01, with a larger difference of .04 for social studies. Therefore, the lower the r the greater the difference with R, since there is more room for improvement. When r is already high, there is only minimal room for such improvement.

The results appear to coincide with the general assumption that multiple-R analyses produce correlations that are only a little higher than the best simple r.

CHAPTER VI

POLICY IMPLICATIONS AND CONCLUSIONS

The main points of the final chapter include an evaluation of the reference tests as potential moderating instruments, the conclusions and policy implications based on the findings from the current study and from previous New Zealand research, and directions for future studies in the area of moderation.

CRITERIA FOR EVALUATING A MODERATING TEST

It was noted earlier (Chapter II), that Gilmore (1979) had identified a list of important crietria to consider, when evaluating the suitability of a moderating test. Most of these criteria add up to major arguments, firstly for the development of a test construct based on developed abilities, and secondly, for the inclusion of a multiple-matrix sampling procedure.

The arguments for (or advantages) associated with tests of developed abilities are that they:

* reflect pupil skills which result from school study and teaching quality;
* possess reasonable face validity to teachers and pupils;
* minimize coaching and backwash effects;

* are essentially syllabus free; and
* do not emphasize rote memorization of facts.

The arguments for (or advantages) associated with a policy of multiple-matrix sampling are that it:

* allows for the coverage of a wide range of curricular and behavioural objectives so that a broad spectrum of pupil ability may be assessed;
* reduces the administration time of the testing programme;
* avoids undue examination stress; and
* minimizes the chance of cheating by pupils

While each of these criteria represent important characteristics of a suitable moderating test, the most crucial criterion must be the test's ability to detect real differences in the pupils (classes or schools). The current study has clearly demonstrated the potential of developed abilities tests, based on predominantly vocabulary and comprehension content, to discriminate successfully between class groups, with a high degree of sensitivity.

POLICY IMPLICATIONS AND CONCLUSIONS

This section will attempt to relate the findings from the current study and those from previous research in terms of practical application, and particularly with regard to policy recommendations presented in the Ross Report (1986).

Reference Test Approach to Moderation

This study has successfully demonstrated the suitability of using reference tests to act as moderating devices. More specifically, it was shown that relatively short reference tests in the common core-subjects of English, mathematics, science, social studies and general aptitude can reliably discriminate between the performance of class groups in the same way as the corresponding School Certificate Examinations; but most importantly, without the associated disadvantages linked to an external examination system.

In a similar way, Gilmore (1979) has also demonstrated the feasibility of using short subject-based reference tests to predict group performance in a similar manner to the School Certificate Examinations. Hulbert (1978) differed somewhat in his approach by using a single reference test - not subject specific - which correlated moderately well with individual School Certificate Examination marks. Again, his findings produced similar trends to those of Gilmore and the present study (using the individual pupil as the unit of analysis). It is likely that if Hulbert had conducted regression analyses based on data at a group level of analysis, he would have produced correlations close to 0.9 with the School Certificate Examinations in his sample.

Overseas, reference-tests have been successfully employed to maintain standards at various levels in Australia, Sweden, and the USA. In particular, the Australian experiences have shown

test-based approaches to be a more reliable and efficient approach than a teacher-based consensus approach, or indeed, no moderation at all. More reliable because it takes the responsibility off the teachers to act as moderators, and therefore, removes the subjective "human" factor from the moderation process. More efficient because such a moderation scheme would leave teachers free to concentrate on teaching and learning. It would be relatively inexpensive once developed on a national basis, time efficient and generally less stressful for pupils.

It has already been noted that one recommendation in the Ross Report (1986) called for research into alternative means of moderation for Sixth Form Certificate, specifying - among other options - the use of reference tests (p. 64). In addition, evidence from Elley's (in progress) survey of teachers on assessment and moderation issues, revealed the use of reference tests to be the second most popular method of moderation, after criterion-based assessment. It was argued earlier, mainly in terms of research, time and resources, that criterion-based assessment is likely to remain merely a policy proposal for some considerable time yet. As a positive interim step, the implementation of a reference test approach to moderation is currently the most realistic recommendation towards total internal assessment of the Fifth and Sixth Form levels in New Zealand.

Although the Ross Report recommendation was for moderation in Form 6, the current study has focused at the School Certificate

level in From 5. This decision was made for three main reasons. Firstly, the School Certificate Examinations provided the most suitable criterion to validate the reference tests against. The current system of moderation for the Sixth Form Certificate (SFC) is based on the previous year's School Certificate Examination results. By demonstrating that the reference tests are able to predict performance on the School Certificate Examinations with a high degree of precision, then reference tests could similarly act as a moderating scheme for SFC, but without those constraints and disadvantages that are inherently linked to an external examination.

Secondly, all previous local studies on the issue of moderation have also focused at the Form 5 level, and by evaluating at the same level it made for an easier comparison of results.

Thirdly, there is no reason to suppose that the correlations would differ substantially between Form 5 and Form 6 pupils. The majority of pupils (approximately 60% and growing) stay on for a Sixth Form year nowadays.

Tests of Developed Abilities

The theoretical construct of developed abilities was shown to be an entirely suitable basis upon which to develop the reference tests. As previously defined, a test of developed abilities measures only those general school-related abilities (e.g. general comprehension, interpretation of and application of

basic concepts and reasoning skills) within a subject area, as distinct from the recall of straight factual knowledge related to a particular syllabus.

Most research with tests of developed abilities has occurred in Australia with the develoment of ASAT (and its earlier version, TEEP) in the form of a state-wide moderating test. Locally, Gilmore (1979) focussed on tests of developed abilities quite extensively, as did the current study, while Hulbert's (1978) EMT also fell into this category, although not specifically described as such.

Clearly, the findings of this study have reconfirmed the potential of the developed abilities construct as a highly sensitive discriminator of class performances on the core School Certificate Examination subjects. If the reference test approach to moderation were to be implemented as policy at either Form 6 or 5 then it would have to be recommended that the subject tests be constructed as developed abilities tests.

In contrast, the scholastic aptitude test - although proving a quite powerful predictor of class performance on the School Certificate Examinations in its own right - lacked face validity. It probably did not reflect teaching quality and was shown to be highly unpopular with teachers as a moderating test (Elley, in progress). However, despite this unpopularity, its technical efficiency - especially as part of the multiple regression equations - would probably warrant its inclusion as

a component of a moderating test battery. Certainly the high intercorrelations between different subject tests, indicates a large general underlying factor in the School Certificate results. If this is an estimate of Spearman's 'g' factor, then it is surely measured more efficiently by tests of verbal intelligence.

Vocabulary and Comprehension (English, Science and Social Studies)

The full potential of a combination of vocabulary and comprehension abilities as content for use in moderating tests, until now, had not been properly assessed. In her study, Gilmore (1979) incorporated only limited prose passage material in two subjects, while Hulbert's (1978) EMT, although a general comprehension test, was not given in conjunction with vocabulary, or as separate subject-based tests.

One reason for a vocabulary/comprehension format arose from a consideration of the Progressive Achievement Test series which are widely used in New Zealand schools. The separate PAT Reading Vocabulary and Comprehension tests, although not subject-specific, were constructed to measure similar sorts of skills as tests of developed abilities. Other research had identified a strong relationship between vocabulary and comprehension. In fact, the relationship is so strong it should be asked as to the necessity of incorporating both components. Why not employ just one or the other?

Logically, there is a difference in their underlying constructs. Vocabulary measures knowledge directly - the outcome of experiencing

much reading and listening, etc. Comprehension focusses more on current processing. The intention has never been to consider vocabulary as a moderating test in its own right, (such a test would over-emphasize the role of knowledge) but rather, as a component of a larger test battery. Vocabulary has several features - for example, it is difficult to coach for; it reflects a pupil's ability to learn quickly (Anderson and Freebody, 1981) - which make it particularly suitable for moderation purposes. By combining it with comprehension, a format with greater flexibility and robustness is developed, designed to measure a wider range of school-related tasks and skills.

A comparison of the separate Vocabulary and Comprehension results showed the correlations of the former (a 12 item test) to be overall slightly higher than for the latter (a 17 item test) with the School Certificate Examinations. In combination, the correlations showed significant improvements. The findings generated from this study support the use and continued research of the vocabulary/comprehension format as suitable test content, not only in the context of moderation, but also as a measure of general achievement.

If a reference test policy were adopted, then the vocabulary/comprehension format must be given serious consideration for inclusion as a component for English, science and social studies subject tests.

Format of Test Items

Survey results from Elley (in progress) and teacher feedback from Hulbert (1978) had revealed a long-standing suspicion about the use of the multiple-choice item, despite its technical efficiencies and popularity with test constructors. With this in mind, the secondary investigation of this study focussed on the performance of alternative item formats - namely, the open-ended, cloze and essay items - in comparison to the technically proven multiple-choice item.

The open-ended format was used with the Vocabulary and Mathematics Tests. This item demands recall (as against recognition for the multiple-choice item) and for pupils to work through and write down their responses. The cloze procedure is used specifically with comprehension tests, and in this situation pupils were required to replace every tenth word using the context of the passage only, to provide cues as to the correct word.

An analysis of the results revealed two main findings. Firstly, for the prediction of School Certificate class means, there was no significant difference in the size of the correlations between the two item formats. Secondly, for the prediction of School Certificate class standard deviations, the multiple-choice tests produced markedly better correlations than the open-ended/ cloze tests.

However, the prediction of the class means is substantially the more important of the two parameters. The basic task implied

by moderation is to maintain standards by being able to successfully discriminate between the ability level of each class (or school) in each major subject. Therefore, determining class means is more crucial for moderation purposes than knowledge about the range of ability in each class, as indicated by the standard deviation score. In addition, because the School Certificate marks undergo scaling, the effect on the standard deviation makes it less predictable. To enhance the prediction of the standard deviation teachers could have some flexibility to specify the range of their pupils' grades.

The cloze procedure, using synonym replacement marking, produced slightly higher reliability estimates and prediction correlations in comparison to exact replacement marking only. However, such a small improvement in the results must be considered in relation to turn-around-time for the test papers and the additional financial resources required to carry out a longer marking process.

In terms of practical application as a nationally developed and administered moderation test, the open-ended item may be considered cumbersome in a context where the emphasis will often be on computer marking for speed and efficiency, and to eliminate errors. However, the cloze procedure should prove highly adaptable to computer marking, especially if scoring is limited to exact replacement only. In relation to the number of items per unit of time, the cloze format is more efficient than the multiple-choice item and must be considered as a valid and reliable alternative format for measuring comprehension of subject-based prose passages.

The Essay Test was included on the grounds that most English teachers feel their subject to be essentially one of communication, and therefore, attempts should be made to assess the quality of written expression. The developers of ASAT have suffered similar criticism and in recent editions an essay writing component has been added.

Despite its modest correlation with School Certificate English in this study and Gilmore's (1979), the essay did contribute a small improvement in the multiple regression equation for the prediction of School Certificate English class means. Although it has inherent problems of marker unreliability, the essay is very important for its contribution to face validity, and must still be considered as an essential component of an English moderating test.

Multiple-Matrix Sampling (MMS)

MMS was developed as a design for measuring group differences; hence its direct application to group moderation. The advantages associated with MMS have already been summarized earlier in this chapter. Suffice it to say, that it lends itself perfectly to the task of moderation, and currently must be considered as the most efficient sampling procedure for use with test-based moderation.

Gilmore's (1979) application of MMS was more extensive than that of the current study, incorporating a specialized MMS parameter estimation computer programme which took into account the particular form of each test responded to by each pupil. However, she also

undertook a supplementary analysis which ignored the multiple forms and simply analyzed the results based on raw scores. The comparison of the two sets of results showed there was no significant difference in the multiple-R's obtained with the two procedures. This demonstrated an important technical aspect, by showing that parameter estimates based on raw scores were just as valid and reliable as those based on more complex MMS programmes. This has crucial implications for practical application, if schools were to become responsible for moderation, either totally or in part.

Thus, in terms of policy implications for moderation, MMS is currently the most appropriate and efficient sampling design available, regardless of which type of parameter estimation programme is used in conjunction with it.

Multiple Regression Predictions vs Simple Correlation Predictions

A comparison of two methods used to generate correlations for the prediction of School Certificate class parameters was conducted in the light of possible policy implications. The first method involved a relatively complex multiple regression programme, while the second method was based on simple correlations using unweighted combinations of predictor variables.

It was shown (Table 5.10) that there were only small differences between the respective sets of correlations. The multiple-R predictions were generally slightly higher than the simple r's.

If future policy were to dictate that schools would become responsible for moderation, either totally or in part, then this technical aspect will have important practical application.

Individual Subject Reference Tests

The current study has focussed on the prediction of School Certificate Examination subjects using corresponding subject-based reference tests. This emphasis on subject independence has also been recommended in the Ross Report (1986, p 15), as a desirable quality of potential moderation schemes.

Gilmore (1979) limited her predictor tests to the three core subjects of English, mathematics and science, predicting a range of School Certificate subjects with a high degree of sensitivity. The current study developed a general purpose Social Studies Test to predict School Certificate Geography, History and/or Economics as a combined Social Studies Examination criterion. The results for Social Studies proved quite satisfactory, although in general, the correlations were lower than those for the other subjects. Prediction of the three individual School Certificate subjects may have produced better results.

Overall, these results appear as reasonably promising, and suggest potential for further development of subject-based reference tests. However, there is clearly much research that needs to be completed, with not only social studies related subjects, but also with so called "non academic" subjects, before any recommendations can be reliably discussed.

Alternative Approaches to Using Reference Test Scores

The lack of research noted about the moderation of non-core subjects may necessitate the development of alternative moderating approaches, if only on an interim basis. Two other possibilities include a total score (or omnibus) test and two-subject moderation in English and mathematics.

The total score or omnibus test consists of components from several core subject areas which are derived to produce a single test score. The single score can then be used to predict a wide range of School Certificate subjects. The current study also evaluated this approach indirectly by combining the four subject-based reference tests into one omnibus test and analysing the total score results. The prediction of class parameters was generally slightly higher than the best optimal multiple-R's for each School Certificate subject (see Tables 5.7 and 5.8).

This approach appears to have some technical merit, as well as producing excellent results. In Australia, ASAT provides a good example of an omnibus test in practical application. If, for whatever reason, policy makers decided to disregard subject-based reference tests, then development of an omnibus test must be given strong consideration.

The second alternative is that of two-subject moderation in English and mathematics. This would involve moderating all other subjects using the two core subjects that are taken by all, or most pupils. This approach has been adopted in New South Wales

(Year 10) and was seriously considered in Ontario.

The current study, and that of Gilmore (1979) have both shown the combined performance of the English and mathematics tests to be very powerful predictors of class parameters across a range of School Certificate subjects. These results are reported below in Table 6.1.

TABLE 6.1

Prediction of School Certificate Examination Class Means Using the English and Mathematics Reference Tests Only.

Predictor Tests	School Certificate				
	English	Mathematics	Science	Social Studies	Total
Chamberlain ('88):					
EV; EC; Ey.	.90	.87	.90	.82	.91
M	.79	.97	.91	.86	.94
EV; EC; Ey; M.	.92	.98	.95	.89	.98
Gilmore (1979):					
English	.94	.82	.82	.77	-
Mathematics	.92	.93	.91	.81	-
Median	.92	.93	.91	.82	.94

All correlations statistically significant at $p < 0.01$.

Clearly, the English and mathematics reference tests, both singly and in combination, are particularly effective predictor

tests across all subject areas. In fact, they correlate more highly with School Certificate Science and Social Studies than the corresponding reference tests do. The combined English and Mathematics Tests produced correlations which were equivalent to those produced by the best optimal predictors reported in Table 5.7.

Despite its statistical effectiveness, this approach is unlikely to be a strong option, especially for use at the higher Sixth Form level where mathematics is not compulsory. It is also likely to lack strong face validity. However, it might have some application as a means of determining and maintaining general standards in the fifth form, once the School Certificate Examinations have been abolished.

Inter-Class vs Inter-School Moderation

The final point of interest concerning practical application is the issue of what size group moderation should focus on, namely should it attempt to discriminate between classes or between schools. The two studies which focussed on group prediction, Gilmore's (1979) and the current study, did so at the class level, mainly because to arrange sufficient numbers of school groups would demand a huge research undertaking. It is also likely that schools would differ less significantly in performance level (i.e. mean score) than individual classes would.

Notwithstanding the lack of research at the school level, Gilmore has identified the main features associated with the two levels of moderation. While inter-school moderation would entail

individual teachers having a greater role in the moderation process in that they would be responsible for maintaining comparability between classes within their school, it also means that each school would have to produce a rank-order of their pupils in each subject. Evidence already exists (e.g. McCausland and Hall, 1985) which shows teachers to be unreliable at discriminating between the ability levels of separate classes. Thus schools may be forced to use their own moderating tests, which in many cases would be technically less than satisfactory.

By focussing moderation at the class level, however, it would become a one-step process with a national assessment body having total responsibility, thus leaving the schools free to utilize all their resources for teaching and learning purposes.

With regard to policy implications, moderation at the inter-class level in conjunction with a nationally developed moderating test and a national assessment body would appear to be a more valid and reliable approach to moderation than the alternative of inter-school moderation.

GENERAL CONCLUDING REMARKS

This study set out with the primary aim of investigating further the potential of a reference test approach for moderating internal assessment at the upper secondary school level. More specifically, the main task was to develop and validate reference tests in four core subject areas, and with the addition of a

general ability test, to assess the sensitivity with which the class parameters of School Certificate Examinations could be predicted from pupil performance on the reference tests. In addition, as a secondary aim, the suitability of alternative question formats was evaluated against the traditional multiple-choice item.

The general development and design of the study was a reflection of overseas research and local studies by Elley and Livingstone (1972), Hulbert (1978) and Gilmore (1979). It also represented a partial replication of Gilmore's study by recognizing the importance of focussing at a group (or class) level of analysis, and incorporating MMS and the developed abilities construct as essential features for moderating through reference tests.

The outcomes of the study successfully demonstrated subject-based reference tests to predict School Certificate Examinations with a high degree of sensitivity. Therefore, they must be considered as eminently suitable for the task of moderating at the Sixth Form level to ensure comparability of national standards. In addition, the alternative item formats were found to perform equally as well, compared to the multiple-choice item, in predicting class ability levels.

The final stage of the study attempted to summarize the findings with those from previous studies, in terms of policy implications and/or practical application of moderating schemes, to enable the introduction of full internal assessment.

This study was undertaken with the knowledge that only a few local studies existed. Yet the issue of external examinations and internal assessment has been a hotly debated issue since the early 1970's, and while many other developed nations have long since introduced partial or total internal assessment in their senior schools, New Zealand has only recently introduced any reforms. Now suddenly, in view of the recommendations of the Ross Report (1986), full internal assessment at the fifth and sixth form level is to be introduced without an adeuqate method of moderation having been conceived.

The majority of teachers, while happy about the introduction of internal assessment, are just as anxious that a satisfactory method of moderation be introduced concurrently. Pupils fear the prospect of being judged by the particular school they attended rather than by what credentials they have attained. Similarly, parents and employers are also affected by the current changes in school assessment procedures.

While there is some overseas research to benefit from, it is equally important to develop our own research, based specifically on the New Zealand situation and requirements. This study should contribute in some measure to the issue of moderation, and more specifically to the reference test approach, at the upper secondary school level. Hopefully, it may stimulate other researchers and more Government funding to promote what is currently a vitally important area of educational research.

Future Studies

Clearly, there is an urgent need for research into all types of moderating procedures with a range of subjects, both academic and practical in nature.

In relation to specific findings from the current study, several aspects would benefit from further attention. These include:

1. a cross-validation to optimize the multiple regression equations;
2. a further investigation of the potential of the Social Studies reference test in predicting separate social studies related subjects (e.g. geography, history);
3. the development and validation of individual reference tests for predicting accounting, foreign languages, separate sciences, etc.;
4. an investigation of the ability of the reference tests to predict other non-academic subjects;
5. an investigation into why the open-ended/cloze item format failed to adequately predict the School Certificate class standard deviation estimates;
6. a further evaluation of alternative item formats in comparison to the multiple-choice question; and
7. a further analysis of reference tests in an omnibus format, and their effectiveness as academic predictors.

The golden rule in life is moderation in all things.

Terence, 166B.C.

We are confident that school assessments, used in conjunction with tests not bound to a specific syllabus, will be effective predictors of success and valuable guides for students in choosing future studies.

Radfort Report, Queensland, 1970

The advantage of using tests, particularly objective tests, as moderating instruments is their technical superiority to other forms of moderation.

J.F. Kerr, 1968

APPENDICES

19 June 1986

APPENDIX A

LETTER OF INVITATION TO SCHOOLS TO PARTICIPATE IN THE EXPERIMENTAL TESTING PROGRAMME

Research on Internal Assessment

We are seeking your assistance in conducting a research study on internal assessment. Glenn Chamberlain, a post-graduate student working under my supervision, is currently undertaking an M.A. research project. He has specialized in educational evaluation, and his topic focuses on moderating procedures for internal assessment at the senior secondary school level. I am sure you will agree that we need to explore all possible avenues of moderation if new policies are to be effective.

Glenn's proposal is to develop two kinds of reference tests that are relatively 'syllabus free', and to compare their ability to predict students' School Certificate marks, in academic subjects. The results should clarify the suitability of the tests for use in moderating fifth form assessments between schools. To do this, he would like access to 16 Form 5 classes across four schools, and we are therefore seeking permission to administer the tests to a cross-section of four classes at your school.

More specifically, we would like approximately two and a half hours of testing time with each class, preferably spread over two different days, and early in the third term. The proposal is to administer tests of aptitude and underlying skills in English, Maths, Science and Social Studies, plus a scholastic aptitude test and an English essay. Two different question types, multiple-choice and open-ended, will be compared, and extensive use will be made of the kinds of syllabus-free tests developed in other places and adapted for use in local schools.

We hope that the project might be helpful for your students, coming as it would before the School Certificate exams. We would naturally treat the results with the utmost confidence, but would provide you with pupils' scores if you would like them. The study should prove beneficial in clarifying some of the policy problems and prospects in moderating by means of reference tests.

We would be most grateful if you and your staff feel able to assist in this way. I will phone you in a few days to discuss it further with you.

Yours sincerely,

APPENDIX B FORM A

MULTIPLE CHOICE

VOCABULARY

TEST

This test attempts to measure your knowledge of common school-related words. The words in this test have been selected from the subject areas of English, Science or Social Studies. The words relating to each subject area have been set out in three separate sections.

INSTRUCTIONS

For each UNDERLINED word, choose the ONE answer which is CLOSEST in meaning to it. CIRCLE the appropriate letter that corresponds to your answer on the separate ANSWER SHEET.

EXAMPLE QUESTION:

The little girl is good at football.	
A	sick
B	mean
C	small
D	large
E	lazy

The word "small" is closest in meaning to the underlined word. Therefore, option C is the correct answer. Thus, a circle is drawn around the letter C, as in the example on the separate answer sheet.

If you want to CHANGE your answer, simply cross out the error and clearly indicate your new choice.

There are 36 questions in total.

Do NOT spend too much time on any one question.

You will have 15 minutes to complete the test.

Please do NOT write in or mark this question booklet.

DO NOT START UNTIL YOU ARE TOLD

ENGLISH VOCABULARY

1. We have already anticipated this conclusion.
   - A supplanted
   - B foreseen
   - C endorsed
   - D opposed
   - E recommended

2. The tourists asked for more time for orientation.
   - A shopping
   - B relaxation
   - C entertaining themselves
   - D finding their way round
   - E taking photographs

3. Consider the derivation of this word.
   - A meaning
   - B origin
   - C spelling
   - D amibiguity
   - E pronunciation

4. The writer's use of a metaphor was most appropriate.
   - A a vivid expression calling one thing by the likeness of another
   - B an expression which conveys its meaning by its sound
   - C a rhythmical piece of prose writing
   - D a well-worn expression in constant use
   - E a phrase consisting of words starting with the same sound

5. The politicians said it was a contentious matter.
   - A a fascinating
   - B a debatable
   - C an unfortunate
   - D a pleasing
   - E an unusual

6. The writer's epilogue was particularly impressive.
   A letter
   B inscription
   C concluding section
   D poem
   E style

7. The tangi lasted for several days.
   A dance festival
   B Maori tournament
   C mourning ceremony
   D storm
   E feast

8. The girl speculated on the teacher's intentions.
   A wondered about
   B approved of
   C was upset by
   D ignored
   E disagreed with

9. The book's appendix contains all the required information.
   A list of book titles consulted
   B additional material
   C list of corrections
   D list of unusual terms
   E subject and page number references

10. The chairman began a lengthy discourse.
   A joke
   B agenda
   C poem
   D criticism
   E speech

11. The scientist was excited by the subsequent reactions.

A following
B unexpected
C powerful
D huge
E constant

12. The sample was rather heterogeneous.

A large
B irregular
C highly intelligent
D unusual
E varied

SCIENCE VOCABULARY

13. Solidification is important in certain industrial processes.

A dispersion
B reduction
C undulation
D crystallization
E extraction

14. The chemist attempted to precipitate the solution.

A mix
B conserve
C condense
D devitalize
E vapourize

15. The two things occurred simultaneously.

A quickly
B slowly
C at the same time
D one after the other
E very rarely

16. This is an anaerobic expereiment.

A alkali-based
B surrounded by air
C protruded inwards
D embryonic
E without oxygen

17. Our measurements showed it to have maximum velocity.

A weight
B speed
C rotation
D height
E acceleration

18. Digestion is an important process in our body's functioning.

A dissolving of food by stomach acids
B breaking down of carbohydrates into starch and glucose
C breaking down of food into smaller units
D breaking down of starch in the presence of water
E breaking down of sugar into the blood stream

19. Hydration experiments were the focus of discussion.

A the addition of water
B conversion to water vapour
C the loss of water in a compound
D bringing water to its boiling point
E distillation of water-soluble substances

20. The fulcrum proved difficult to produce.

A frame
B pyrite
C alloy
D anion
E pivot

21. The inertia effect was unexpected.

A capacity to do work
B force per unit area
C conservation of energy
D ability of a body to resist acceleration
E force of an object moving at a constant speed

2. The doctor stressed the need for a balanced diet.

A carbohydrates, fats and proteins

B a balance of liquids and roughage in well cooked meals

C the correct number of calories for each individual person

D nutrients in the correct amounts for the body

E a diet excluding red meat, fats, dairy products and sugar

3. Endothermic reactions...

A contain heat

B absorb heat

C reflect heat

D transmit heat

E involve no heat

4. Magnetism has several practical uses.

A a property that gives power to attract

B flow of energy from atom to atom

C a measure of potential energy

D a characteristic of metals at high temperature

E melted magnesium

SOCIAL STUDIES VOCABULARY

5. These are our elite members.

A top ranking

B revolutionary

C new

D highly educated

E conservative

6. Public revenue.

A transport

B service

C works

D expenses

E income

7. Allies can provide numerous benefits.

A narrow streets in old cities

B countries in partnership

C voting blocs in urban areas

D gradual absorption of minority languages

28. Depreciation must be taken into account.

A economic weakness
B critical written review
C without hope of recovery
D lowering of value
E downturn in market consumption

29. The institution could not be supported.

A government
B referendum
C official statement
D law
E agency

30. Ethnocentrism was a common factor of their society.

A belief in the superiority of one's own group
B physical separation of different ethnic groups
C language differences in a social hierarchy
D maintenance of one's ethnic identity over several generations
E a trend towards single-mindedness

31. Topography was a central theme at the conference.

A recording of prehistoric fossils
B study of the social enviroment
C discription of physical features
D study of mountains and plateaus
E the analysis of top-soil

32. The value of socialism was discussed by the politicians.

A private ownership of the means of production
B political ideology of voluntarism
C belief in state ownership
D control of the economy by a few
E increased wages and a major reduction in working hours

33. One of our basic commodities.

A goods
B requirements
C industries
D privileges
E traditions

34. Subsistence problems are now levelling off.

A social problems in an urban community
B the ability to manage money
C a means of work to meet basic living needs
D the collapse of unstable ground
E a stable family environment

35. The weathering effect has been variable.

A a form of radiation on plant life
B the distribution of minerals in the top-soil
C a reduction of matter into energy
D the action of air and rain on the physical terrain
E the effect of insufficent rain on agricultural crops

36. Does this reveal your doctrine?

A discipline
B diversity
C strategy
D medical opinions
E belief system

IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

NAME(please print):_____ APPENDIX C _____

AGE(in years):_____ SEX(circle one): M F

FORM CLASS:_____ SCHOOL:_____

FORM A

OPEN - ENDED
VOCABULARY
TEST

This test attempts to measure your knowledge of common school-related words. The words have been selected from the subject areas of English, Science and Social Studies. The words relating to each subject area have been set out in three separate sections.

INSTRUCTIONS

For each UNDERLINED word, respond by writing a single WORD or PHRASE (do NOT write a complete sentence) which has the SAME or nearly the same meaning to it. WRITE your answer in the QUESTION BOOKLET in the space provided.

EXAMPLE QUESTION:

The little boy is good at sewing.
"small"
"the opposite of big"

In the example, two different types of answer have been provided. The word "small" and the phrase "the opposite of big" both have a meaning which is similar to the underlined word "little". REMEMBER that either of these answers is suitable, you do NOT have to give both.

If you want to CHANGE an answer, simply cross out the error and clearly write in your new answer.

There are 36 questions in total.

Do NOT spend too much time on any one question.

You will have 15 minutes to complete the test.

DO NOT START UNTIL YOU ARE TOLD

ENGLISH VOCABULARY

1. The film director used a fade in the final shot.

2. The teacher began a résumé of the week's work.

3. The legitimate solution is not always the easiest.

4. There was increased interaction between the two countries.

5. Newspapers may well become redundant.

6. The teacher explained what a cliché was.

7. The elder was careful to preserve his mana.

8. The editor discouraged the use of pseudonyms.

9. The man used elaborate paraphernalia.

10. There was little uniformity in the students' responses.

11. We need someone to arbitrate.

12. The test results were ambiguous.

SCIENCE VOCABULARY

13. The light rays were converging.

14. The efficiency of this process is valuable.

15. The class experiment involved distillation.

16. Vapourization involves a change in physical state.

17. Some students were studying the fundamental laws of physics.

18. Catalysts are important in many commercial situations.

19. How much energy is being produced?

20. The sick man's respiration was abnormal.

21. The report focussed on ecology.

22. Chlorophyll is an important substance for plants.

23. The scientist was investigating the molecule.

24. The scientist measured the amount of radiation.

SOCIAL STUDIES VOCABULARY

25. The distribution of the firm's products is increasing.

26. They studied the data comming from the seismograph.

27. The infra-structure of the area was a subject of debate.

28. The problems of anarchy have been well documented.

29. Please respect the local ritual.

30. We made inquiries about the local precipitation.

31. Sovereignty was an important issue to the locals.

32. Inflation will be an important election topic.

33. Such statements could make these people feel alienated.

34. The sociologist analysed the social stratification of the group.

35. We plan to petition our leader.

36. We demand secular education for our children.

IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

MULTIPLE CHOICE

COMPREHENSION

TEST

This test attempts to measure your understanding of a series of prose passages. Some of the questions will involve the interpretation of graphs, diagrams or maps. All of the passages have some connection with English, Science or Social Studies themes. There are a total of 9 passages grouped according to the three subject areas.

INSTRUCTIONS

Briefly read each passage right through. Then answer the following questions by referring back to the passage as often as you need to.

Please respond by CIRCLING the appropriate letter on the separate ANSWER SHEET to the ONE option which you consider to be the BEST answer.

EXAMPLE QUESTION:

How many hours are there in one day?

A 12
B 23
C 24
D 48
E None of the above

Since there are in fact 24 hours in one day, then option C is the correct answer. Thus, a CIRCLE has been drawn around the letter C in the example on the separate ANSWER SHEET.

If you want to CHANGE your answer simply cross out the error and clearly indicate your new choice.

There are 51 questions in total.

Do NOT spend too much time on any one question.

You will have 50 minutes to complete the test.

Please do NOT write in or mark this question booklet.

DO NOT START UNTIL YOU ARE TOLD

ENGLISH COMPREHENSION

PASSAGE 1.

Answering employment ads is not only a matter of industry and persistence; it is also a matter of skill. I advise two simple rules: You must present all of your qualifications, regardless of whether or not they are asked for, and you must clearly bring out the particular qualifications the employer specifies in his ad.

(9) Sending a resume or general statement of all of your qualifications readily satisfies
(11) the first rule. To take care of the second, the resume should be accompanied by a letter specifically referring to the qualifications requested in the ad you are answering.

(15) Here are a few important don't$:

(16) Don't simply put a copy of the resume into an envelope and mail it off without a letter. It is not good manners. It indicates no strong desire for the particular position, but rather an "I'm here if you want me" attitude. It requires the recipient to locate in the resume the qualification he has asked for. It may, if the firm is running more than one ad, be misunderstood as an application for the wrong position. Most important, it neglects your only chance to point out the special qualifications you may have.

(29) Don't use a mechanically reproduced form letter.

Don't write a perfunctory and stilted letter of the sort employers receive by the dozens, such as:

"In answer to your ad, the enclosed resume shows I am very well qualified for the position. I will be glad to hear from you with regard to an interview."

This type of letter accomplishes exactly nothing - because it says nothing.

1. The word "resume", as used in the passage refers to

   A an accompanying letter

   B a list of important "don'ts"

   C a summary of qualifications

   D a carbon copy of a letter of application

   E an explanation of your reasons for applying

2. A good letter of application should

   A indicate that the writer is eager to get the job

   B point out the writer's special qualifications for the job

   C mention the particular position advertised

   D refer to the relevant facts in the writer's resume

   E do all of these things

3. The passage falls into two distinct sections. The second section begins with

   A "Sending a resume ..." (line 9)

   B "To take care of the second ..." (line 11)

   C "Here are a few important don'ts: ..." (line 15)

   D "Don't simply put a copy ..." (line 16)

   E "Don't use a mechanically ..." (line 29)

4. The author's purpose in this passage is to teach the reader how to

   A write a satisfactory business letter

   B follow formal instructions

   C secure an interview with a businessman

   D assess his/her own qualifications for a job

   E apply by mail for an advertised position

5. The author does NOT include

   A directions for the proper form of a business letter

   B advice on being considerate to the busy executive

   C the suggestion that the applicant state all his/her qualifications

   D advice on how to make a good impression

   E a list of things that should not be done

6. A person who followed the advice entailed in the passage's last two recommendations would

A write a letter with poor grammatical construction

B write an informal but enthusiastic letter

C write a brief, business-like letter

D write a personalized, tailor-made letter

E avoid writing a letter at all

PASSAGE 2

THE MARKET RESEARCHER

Some wounded seabird seeking shelter
from a storm could not have looked more
forlorn as she stood on our verandah
dripping wet, satchel under her wing.

Doing a survey on something, wanting us
to fill out a questionnaire. We oblige
and begin studying "Designs for Living"
spread across the floor for us to choose

the house we most prefer, size and shape
of rooms, appliances, the how, why and where.
Over coffee we reverse roles, asking her
about the why and who. She turns out to be

another domestic refugee, one of the lucky
ones who escaped into a school hours job
where she meets people, "all kinds of strange
interesting people." But home life is hell.

He drinks, never takes her out, the kids
are uncontrollable, neighbours not really
compatible, no fun, nothing. She packs up
her folios, then pauses. "Funny," she says,
"there's no questions about that in this."

Barry Southam

7. The poet used the "seabird" simile in the first lines because the market researcher

   A was clearly in pain

   B was soaked to the skin

   C had an arm like a wing and a face like a wounded bird's

   D looked wet and miserable

   E she seemed to be like a refugee

8. The poet's use of "doing a survey on something..." and "we oblige ..." imply that he was

   A interested in the researcher's questions

   B required by law to complete the questionnaire

   C co-operating only out of a sense of duty

   D co-operating as a means of gaining an insight into the researcher as a person

   E tired at having to complete yet another questionnaire

9. The market researcher is "another domestic refugee" because

   A her job enables her to escape the problems associated with her home situation

   B it was the only job she could find that allows time for her domestic duties as well

   C her job involves interviewing house - husbands/wives

   D she feels superior while working with people in a situation she is familiar with

   E she detests having to do domestic related work

10. In the poem's final line "there's no question about <u>that</u> in this", what does 'that' refer to?

   A the questionnaire

   B her husband

   C being a domestic refugee

   D life as a market researcher

   E uncontrollable children

11. All of the following are suggested by the poem <u>EXCEPT</u>

   A we should be co-operative with people who call

   B people who conduct surveys are a nuisance

   C many people have unhappy homes and need sympathy

   D people are often different from what they appear to be at first glance

   E a job is often more than just a source of income

PASSAGE 3.

Of all the thousands of words that have been written and spoken about the Chernobyl nuclear power plant disaster in the Soviet Union, the most telling have come from its leader, Mr. Gorbachov. He said what happened showed again what an abyss will open if nuclear war befalls mankind.

The world has indeed been given a most graphic warning. There was a relatively small explosion in the Chernobyl power station and it took place in a reactor encased in steel and concrete. Even so, at least seven are dead, hundreds are in hospital and thousands have had to be evacuated.

The effects and consequences are infinitesimal compared with what would have happened had a nuclear weapon been exploded above ground level. The magnitude of the danger is almost totally beyond human comprehension.

As the Royal Society of New Zealand observed last year we all know the difference between zero and 100 deg. but what do we make of a temperature of 15 million degrees generated by a "small" atomic explosion? How can we really appreciate the destructive capacity of a single modern nuclear bomb several times more powerful than that dropped on Hiroshima?

The stockpile of nuclear weapons distributed around the world contains the explosive power of more than a million Hiroshima bombs. They pose an unprecedented threat to humanity yet somehow the world lives with it. Perhaps the potential of destruction is so awesome we cannot even begin to comprehend it.

An explosion in a nuclear power station is something we can understand, and worry about. At Chernobyl only a little of the nuclear genie escaped from the bottle by accident. When we see the fearful results we can begin to understand what would have happened if it had been nuclear weapons that exploded.

Hopefully the lesson from Chernobyl will not be lost on all the peoples of the world and they will press even harder for nuclear disarmament. The arms race is still being pursued to the point of madness. Now is the time to end it.

12. The word "graphic" (paragraph two) means what?

A diagrammatical

B vivid

C unfortunate

D lucky

E essential

13. The word "small" (paragraph four) has been emphasized by quotation marks because

A the editor was unsure of the exact size of atomic explosions

B it would add more punch to the passage

C scientists cannot estimate the likely effects from large explosions

D small atomic weapons are rarely used nowadays

E the editor wanted to contrast it with the consequences of a large atomic explosion

14. How many nuclear bombs of the type dropped in 1945 are equivalent to the current nuclear stockpile?

A several

B a hundred

C a thousand

D a million

E greater than a million

15. The word "they" (paragraph five) refers to

A the political leaders of the U.S.A. and Russia

B the discoveries of science

C nuclear accidents similar to Chernobyl

D countries involved in the nuclear arms race

E nuclear armaments

16. All of the following are mentioned in the editorial EXCEPT

A the first bomb was dropped on Hiroshima

B an increase in the destructive power of nuclear weapons

C the dangers associated with nuclear power generators

D the need for people to take responsibility for their actions

E the Chernobyl disaster is an important lesson to everyone

17. What is the main purpose of the passage?

A To warn people about the danger of nuclear power accidents.

B To complain about the Russian accident at Chernobyl.

C To show how close mankind was to extinction at Chernobyl.

D To emphasize the urgency of nuclear disarmament.

E To stress the waste of words written about Chernobyl.

SCIENCE COMPREHENSION

PASSAGE 4.

In different environments different kinds of animal communities exist, and in each of these are interrelationships based upon the food habits of the various species.

In a pond or small lake, for example, certain microscopic animals (protozoa) feed upon bacteria, microscopic plants (algae), and other, usually smaller, protozoa. The protozoa, in turn, are fed upon by small many-celled animals, such as rotifers and small crustacean relatives of the crayfish and the crab. These in turn, may be fed upon by larger crustaceans, by water-living insects, and by small fishes. And these, in turn, may be fed upon by still larger fishes. Fish-eating birds may also be involved in this group of relationships of the hunter and the hunted. On land the small herbivores (mouse, squirrel, rabbit) are fed upon by the small carnivores (owl, hawk, fox), while the larger herbivores (cattle, sheep, deer) are preyed upon by the larger carnivores (wolf, cougar, lion).

Such a pattern of interrelationships among animals is called a food chain. Each food chain, whether composed of many links or only a few, is based upon the principles that the predator is usually larger than its prey, and that the lowest link is made up of herbivorous animals.

It is obvious that the smallest animals in a stable food chain relationship must be the most numerous, and that the number of animals in each successive link of the chain must be fewer and fewer. A small pond could support only a few fishes, and it might require many such ponds to support a single fish-eating bird. Also in keeping with these relationships are the generally more rapid reproductive and growth rates of the smaller species.

18. If herbivores represent the first link of a food chain, then the base of the food chain will consist of

    A fishes

    B plants

    C small carnivores

    D large carnivores

    E omnivores

19. Which of the following is true?

A There are more carnivores than herbivores.

B There are about as many carnivores as herbivores.

C There are fewer carnivores than herbivores.

D There are no stable relationships between the number of carnivores and the number of herbivores.

E The number of omnivores is the most crucial aspect of a food chain.

20. Populations of animals in nature may rise and fall successively. Trapping records of the Hudson's Bay Company indicate that fox and rabbit cycles in Canada are of about the same length (7 - 12 years). Which of the following relationships of fox and rabbit populations would be most likely?

A Increases in rabbit and fox populations occur at the same time.

B Increases in rabbit and fox populations occur at different times.

C Increases in rabbit populations come before similar increases in fox populations.

D Increases in rabbit populations occur during periods when fox populations are stable.

E Fox and rabbit populations show no casual relationships to each other.

21. Why are individual tigers found in widely separated areas?

A Tigers tend not to get on well with each other.

B Wide separation is good protection against predators.

C It is difficult for more than one tiger to find adequate shelter in the same place.

D A large animal like the tiger must be at the end of an extensive food chain.

E Advancing city limits have caused a decrease in their numbers.

22. What does the writer try and do in the last paragraph?

A Show there is a relationship between the number of animals and their level in the food chain.

B Demonstrate that growth rate varies directly with animal size.

C Demonstrate that habitat varies directly with animal size.

D Show that water-living and air-living forms can be involved in the same food chain.

E Show that reproductive rates vary for smaller animals only.

23. The word "principles" (paragraph three) as used in the passage means

A probability

B assumptions

C chance

D priority

E issue

PASSAGE 5

Those of us who use hard coal as fuel in our furnaces can learn much while stoking the fire.

First, we can notice that when coal is put into the furnace and the bottom draught is opened, a brisk yellow blaze appears above the coals. Sparks leap upward and vanish. The luminosity of the blaze is mostly due to glowing particles of coal.

After the fire has a good start, it is customary to close the bottom draught and to open a small furnace door. In this way the fire is kept from burning too fast. In a little while a blue blaze can be seen, not burning steadily in any one place, but disappearing and then reappearing with a gentle pop or puff. There is no yellow flame now, and there are no sparks flying upward.

What is the explanation of these simple facts? When carbon, or coal, burns in a good draught the following reaction takes place:

$$C + O_2 \longrightarrow CO_2.$$

That is, one atom of carbon unites with two atoms of oxygen to form one molecule of a gas known as carbon dioxide. This gas is not inflammable, has no smell, and is harmless under most circumstances. It passes up the chimney and is lost in the air.

When the bottom draught is closed, very little air comes up through the layers of coals. Under these conditions, instead of two atoms of oxygen for each atom of carbon, the reaction requires only one. The process is

$$C + O \longrightarrow CO.$$

Since there are two oxygen atoms to a molecule, the chemist would rather write this

$$2C + O_2 \longrightarrow 2CO.$$

With the draught closed, then one atom of carbon unites with one atom of oxygen to form one molecule of carbon monoxide, or CO. Carbon monoxide is a poisonous gas without smell or taste, invisible, and therefore especially dangerous. It would be foolhardy to close the bottom draught but for the fact the carbon monoxide is inflammable. On top of the coals the carbon monoxide burns to the comparatively harmless carbon dioxide. It burns on top of the coals only because plenty of oxygen coming through the open door can reach it there.

The flame flickers because, after burning for a moment, all the CO is gone from that one spot, and the flame goes out. Then the slight draught moving upward brings some more CO; and when the open air is reached, the blue flame appears once more.

24. When the bottom draught is open, most of the light from the fire is caused by

A flying sparks

B the burning of carbon monoxide

C incandescent carbon dioxide

D the lightness from the open draught

E glowing bits of coal

25. The two gases formed in the fire differ in

A odour

B colour

C flammability

D the kind of elements of which they are composed

E pressure

26. Why might it be foolhardy to close the bottom draught in this reaction, if CO were not inflammable?

A CO is poisonous

B CO would burn out of control

C Too much of the gas would be wasted

D CO_2 is inflammable

E CO produces an unpleasant smell

27. Why does the carbon monoxide <u>not</u> burn below the upper surface of the coals?

A Temperature conditions are not suitable for such a reaction to take place.

B Carbon monoxide does not form in sufficient quantity below the top surface of the coals.

C There is too much oxygen for such a reaction to take place.

D Carbon monoxide does not remain below the top surface of the coals long enough to combine with oxygen.

E There is insufficient oxygen for such a reaction to take place.

28. What causes the "flickering" of the blue flames?

A The oxygen supply varies irregularly.

B The reaction rate of CO with O_2 changes rapidly.

C The CO burns faster than it is supplied at any one spot.

D The temperature at the upper surface of the coals fluctuates widely.

E The bottom draught has been closed off.

29. What is the writer's purpose in this passage?

A To describe the properties of the oxides of carbon.

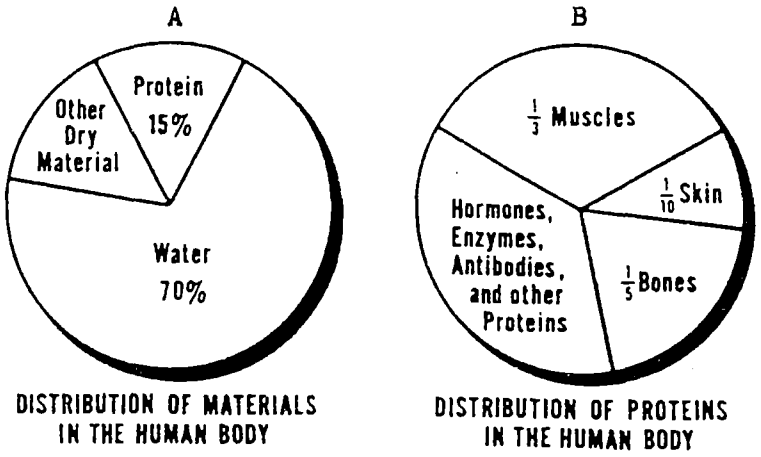
B To tell what causes the blue blaze from coal.

C To demonstrate the danger of operating a furnace without sufficient draught.

D To explain chemical reactions using coal burning as an example.

E To explain the chemistry of the burning of coal in a furnace.

PASSAGE 6.



DISTRIBUTION OF MATERIALS IN THE HUMAN BODY

DISTRIBUTION OF PROTEINS IN THE HUMAN BODY

NOTE: Graph B is an"expanded" view of the protein section in graph A.

30. The human body, according to the data furnished by the two graphs, is composed mainly of

A proteins

B hormones, enzymes, antibodies and other proteins

C muscles

D solids

E water

31. A person weighing 80kg would, according to these graphs, be composed of water weighing

A 8kg

B 15kg

C 56kg

D 60kg

E 70kg

32. In the human body, the distribution of material other than water and protein is equal to (in its simplest form)

A 85/100

B 3/20

C 1/15

D 1/5

E None of the above

33. How many degrees of the circle should be used to represent the distribution of protein? (NOTE: there are 360^{o} in a circle.)

A 15

B 27

C 54

D 60

E 90

34. The ratio of the distribution of proteins in muscle to the distribution of proteins in skin is (in its simplest form)

A 1:3

B 3:1

C $3\frac{1}{3}$:1

D 10:1

E 30:1

SOCIAL STUDIES COMPREHENSION

PASSAGE 7.

The natural environment, commonly termed the land, is the agriculturalist's basic resource. The agriculturalist has the choice of using this resource in two fundamental ways: he may regard the land as a fund type resource to be exploited before moving to another area; or, as a flow type resource to be utilized in such a way as to maintain its productivity indefinitely. The inherent qualities of the land, the pressure of population on the land, and the degree of technological skill and cultural background of the agriculturalist are some of the important factors affecting his response to this choice. To produce goods, in each case, the agriculturalist applies labour and capital to the land in varying proportions and intensities.

The products derived from his plants and animals provide the agriculturalist with his means of livelihood. These products may be consumed directly, by the agriculturalist and his family, or indirectly after some form of processing to give the original product form utility, i.e. make the good into a more acceptable form for consumption. A large volume of the world's agricultural products are not consumed by those agriculturalists who produce them but are exchanged or sold to non-agriculturalists. These products have to be transported, and perhaps stored and processed before consumption. Transport adds place utility, that is, the product is made more acceptable by transfer to a location where it is demanded, and storage may add time utility by providing the good when it is wanted. Agricultural products which are processed, transported, or stored, gain value well beyond that which the agriculturalist receives for them because the cost of utility are added to the original product.

35. Which of the following products has been given the highest form utility?

    A Wheat grains

    B Bread

    C Flour

    D Wheat stalk

    E Wheat seeds

36. If an agriculturalist loses his/her means of livelihood, we conclude that

    A he/she consumes less than he/she produces

    B his/her farm is not productive

    C his/her goods have gained in cost utility

    D he/she has changed his/her occupation

    E his/her goods have decreased in form utility

37. Which of the following is the best example of land being used as a fund type resource?

A Regeneration timber milling

B Grazing of cattle

C Sugar cane plantations

D Gold mining

E Vegetable growing

38. If an agriculturalist chooses to use his/her land as a flow type resource, this means that he/she plans to

A keep his/her land for productive purposes

B move on to new land

C obtain some financial resources

D irrigate the land

E change his/her agricultural policy

39. If a product is consumed directly, this means that it is

A sold as it is

B eaten immediately

C used without processing

D stored till required

E wasted needlessly

40. The costs of utility in the last line, refer to such expenses as those involved in

A growing plants and vegetables

B obtaining labour to harvest the produce

C acquiring and processing land

D treating goods for consumption

E paying for electricity on the farm

PASSAGE 8

To discuss criminology in the Third World in isolation from the fundamental reality of imperialism would be fruitless and totally inadequate and misleading. My argument is that crime is not an autonomous concept which applies to certain kinds of behaviour in all societies and cannot therefore be studied in itself. Rather it is socially determined, essentially by the powerful groups in society for their own purposes; therefore anyone wishing to consider seriously the concept of crime must analyse the following: (a) what groups have power in society; (b) the purpose for which that power is exercised in relation to the designation of behaviour as criminal. Such an analysis will indicate that by and large a minority with power in society determine which acts shall be considered criminal.

In the case of colonial countries, the power lay with the colonial administrators and they exercised it partly through the legal system, primarily against the relatively powerless indigenous population. There can be no serious argument against the view that the fundamental purpose of the colonial legal system was to repress the indigenous population in order to allow the colonial link to be maintained, for whatever economic, strategic (or combination of) reasons.

And I have recently suggested elsewhere that the study of crime in colonial Africa was part of a wider intellectual endeavour which was intended to assist the colonial administration in maintaining its dominance. Numerous writers have argued recently that criminology serves a similar purpose in capitalist countries, that is, it functions most importantly to aid in maintaining the status quo, by providing the justification for repression of the working class both physically and ideologically.

41. The writer believes that criminal behaviour is

A essential to the position of imperialists

B influenced by capitalist forces

C determined by the dominant group

D an autonomous concept

E limited to Third World countries

42. What does it mean to maintain the status quo, in the last sentence?

A To keep certain classes in their place.

B To carry on a good living.

C To uphold law and order.

D To help the cause of capitalism.

E To reduce the chance of violence.

43. The writer believes that criminology was studied in Africa largely in order to

A reduce the crime rate

B discover the universal causes of crime

C help the colonialists retain power

D provide intellectual stimulation for administrators

E make it's societies more civilized

44. What was the writer's attitude towards imperialists?

A He/she adopted an unbiased stance

B He/she disapproved of them completely

C He/she admired their use of the legal system for their own ends.

D He/she thought they were administratively inefficient

E He/she felt they were justified in their actions

45. The indigenous population in this passage refers to

A imperialists

B colonial administrators

C powerful local administrators

D the criminal section of society

E people of the local culture

46. Why does the author think that crime "cannot be studied in itself"?

A A lack of trained criminologists.

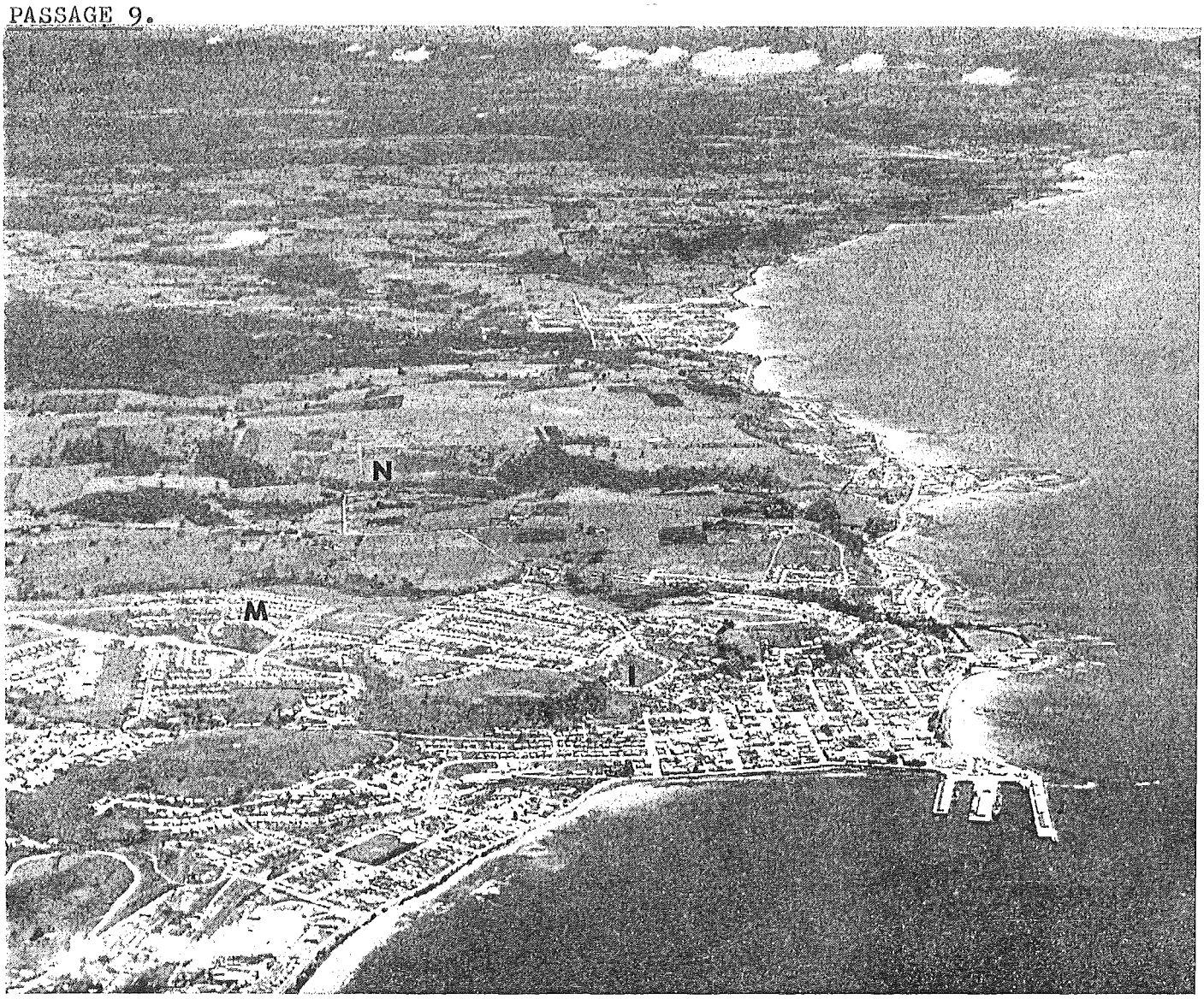
B Criminology is not an exact science.

C The definition of crime differs from one society to another.

D The concept of crime is autonomous in each society.

E Each crime is a unique and autonomous act.

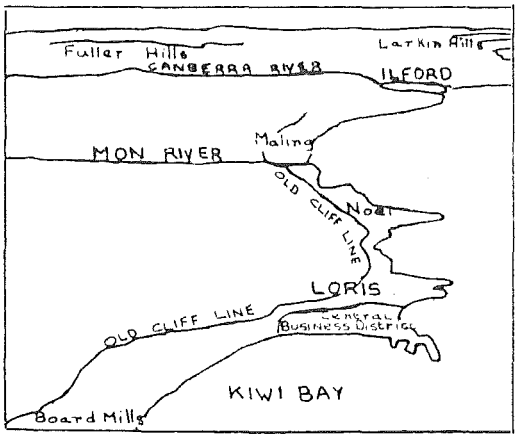
PASSAGE 9.



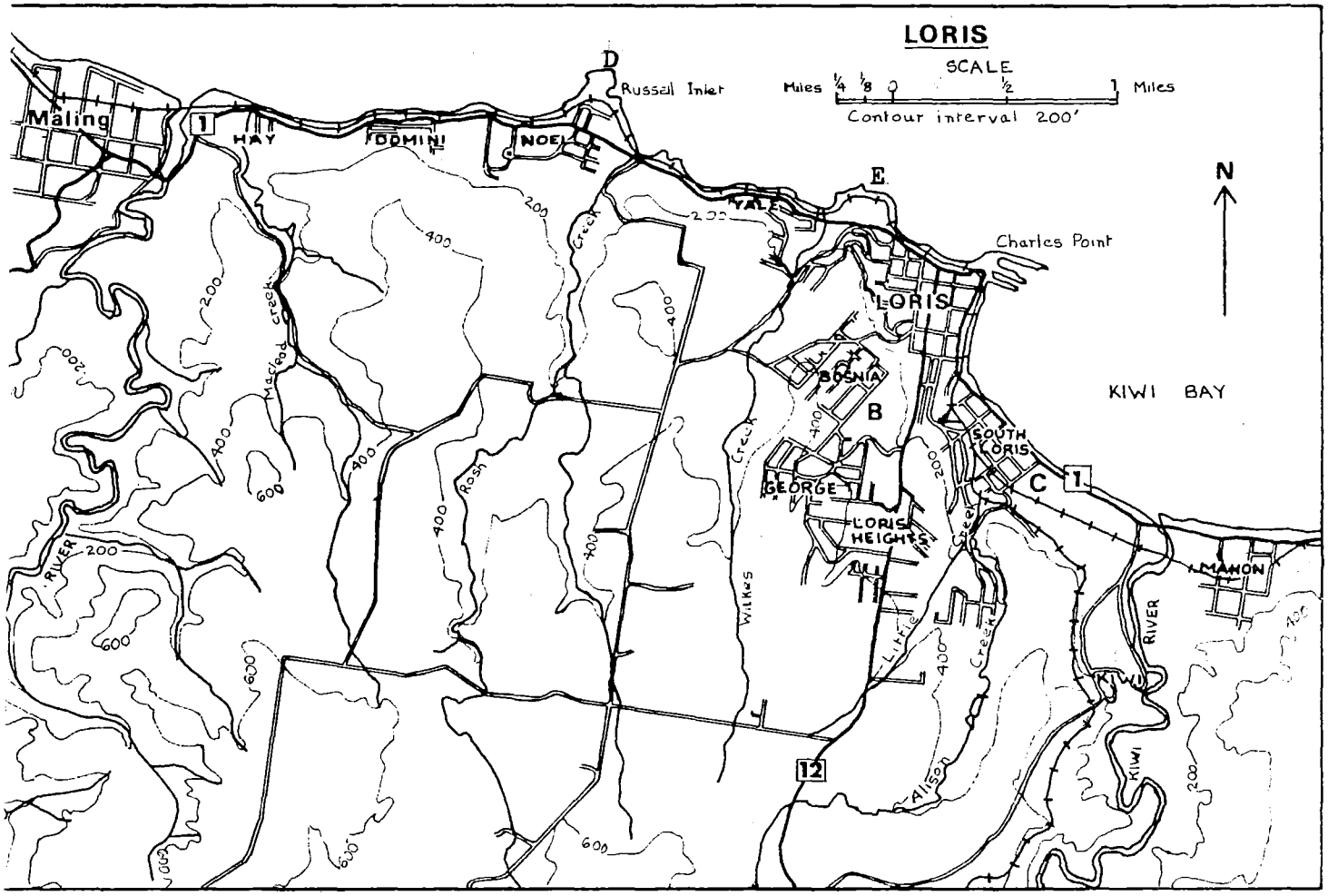
Loris and the North-west Coast

The photograph looks westward and shows a considerable section of the northern coastland west of Loris. The soils on the plateau are used for dairy farming and the growing of cash crops, notably potatoes. Although the plateau slopes more or less gently to the sea, an old abandoned sea cliff occurs some distance behind the present shore along the greater part of the coast. As well as being an important service centre, Loris has developed as an industrial town. A major industry concerned with the manufacture and processing of wood pulp is prominent in the bottom left hand corner of the photograph. Since the photograph was taken, the port has been extended by the construction of new breakwaters protecting the harbour area from the north-east.

Sketch of the Area Shown in the Photograph



Map of Loris



END

) to 25,000 . . . . . SLADE	Railways . . . . .
to 5,000 . . . . . Dorset	Route No. 3
; than 1,500 . . . . . PEARCE	Roads . . . . .
	Creeks . . . . .
	Dams . . . . .
	Contour lines 200

TES

tours are lines on a map joining places of the same height above sea level. base (o metres) is sea level. The constant difference in height between one and the next is known as the contour interval.

47. If a man left the township of Noel and drove eastward along Route 1 for two miles he would be in

A Yale

B Loris

C Domini

D Hay

E Maling

48. Locate on the map the site marked "I" in the photograph. The site marked "I" is most probably

A a clump of trees

B **a slum area beside a creek**

C suburban housing

D a grassy slope

E **a playground**

49. The settlement pattern of the inland area between Wilkes Creek and the Mon River is best described as consisting of

A small townships surrounded by scrub-land

B farms

C small townships in forest clearings

D urban blocks

E forests

50. A train follows the Kiwi River valley after leaving South Loris. While in the area shown in the map and photograph, it would most probably

A climb the old cliff then descend steadily into the river valley

B wind through heavily timbered mountain country

C travel in a southeast direction towards Mahon

D climb the old cliff then travel through heavy bush

E cross over a major highway route

51. From the information on the map we can conclude that the population of Bosnia is

A the same as Mahon

B larger than South Loris

C the same as Loris

D larger than Hay

E greater than Maling

IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

NAME (please print): APPENDIX E

AGE (in years): _____ SEX (circle one): M F

FORM CLASS: _____ SCHOOL: _____

FORM A

OPEN-ENDED/CLOZE COMPREHENSION TEST

This test attempts to measure your understanding of a series of prose passages. Some of the passages will involve the interpretation of graphs, diagrams or maps. All of the passages have some connection with English, Science or Social Studies themes. There are a total of 9 passages grouped according to the three subject areas.

INSTRUCTIONS

FOR PASSAGES 2, 6 AND 9:

Briefly read each passage right through and then answer the questions that follow by referring back to the passage as often as you need to. Respond by WRITING your answer in the QUESTION BOOKLET in the space provided.

FOR ALL OTHER PASSAGES:

These passages have had every 10th word omitted and replaced by a standard length gap. Thus, no matter how small or large the omitted word was, it has been replaced by a gap this long: "_____". You are to respond by writing the ONE word which you think would correctly fill the gap. WRITE your answer in the QUESTION BOOKLET beside the question number which corresponds to the number in the gap.

EXAMPLE QUESTION:

The sun is shining __1__ today. The forecast is __2__ a fine, warm day __3__ an expected high of 20°C __4__ just a light sea __5__.

1. brightly
2. for
3. with
4. and
5. breeze

This is how you are to answer all the passages EXCEPT for passages 2, 6 and 9.

If you want to CHANGE an answer, simply cross out the error and clearly write in your new answer.

Do NOT spend too much time on any one question.

You will have 50 minutes to complete the test.

DO NOT START UNTIL YOU ARE TOLD

ENGLISH COMPREHENSION

PASSAGE 1.

Answering employment ads is not only a matter of industry and persistence; it is also a matter of skill. I advise two simple rules: You must present all of your qualifications, regardless of whether or not they are asked for, and you must clearly bring out the particular qualifications the employer specifies in his ad.

Sending a résumé or general statement of all of ___1___ qualifications readily satisfies the first rule. To take care ___2___ the second, the résumé should be accompanied by a ___3___ specifically referring to the qualifications requested in the ad ___4___ are answering.

Here are a few important don'ts:

Don't ___5___ put a copy of the résumé into an envelope ___6___ mail it off without a letter. It is not ___7___ manners. It indicates no strong desire for the particular ___8___, but rather an "I'm here if you want me" ___9___. It requires the recipient to locate in the résumé ___10___ qualification he has asked for. It may, if the ___11___ is running more than one ad, be misunderstood as ___12___ application for the wrong position. Most important, it neglects ___13___ only chance to point out the special qualifications you ___14___ have.

Don't use a mechanically reproduced form letter.

Don't ___15___ a perfunctory and stilted letter of the sort employers ___16___ by the dozens, such as:

"In answer to your ___17___ the enclosed résumé shows I am very well qualified ___18___ the position. I will be glad to hear from ___19___ with regard to an interview."

This type of letter ___20___ exactly nothing - because it says nothing.

Write your answers next to the question numbers (below) which correspondto the numbered gaps in the passage.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

13.

14.

15.

16.

17.

18.

19.

20.

PASSAGE 2.

THE MARKET RESEARCHER

Some wounded seabird seeking shelter
from a storm could not have looked more
forlorn as she stood on our verandah
dripping wet, satchel under her wing.

Doing a survey on something, wanting us
to fill out a questionnaire. We oblige
and begin studying "Designs For Living"
spread across the floor for us to choose,

the house we most prefer, size and shape
of rooms, appliances, the how, why and where.
Over coffee we reverse roles, asking her
about the why and who. She turns out to be

another domestic refugee, one of the lucky
ones who escaped into a school hours job
where she meets people, "all kinds of strange
interesting people." But home life is hell.

He drinks, never takes her out, the kids
are uncontrollable, neighbours not really
compatible, no fun, nothing. She packs up
her folios, then pauses. "Funny," she says,
"there's no questions about that in this."

Barry Southam

21. What was it about the market researcher that caused the poet to use the "seabird" simile in the first few lines?

22. What does the poet's use of "doing a survey on something ..." and "we oblige ..." imply?

23. Why is the market researcher "another domestic refugee"?

24. In the poem's final line "there's no question about <u>that</u> in this", what does '<u>that</u>' refer to?

25. There are several important themes that can be identified with the "The Market Researcher" poem. Name <u>ONE</u> of these themes?

PASSAGE 3.

Of all the thousands of words that have been written and spoken about the Chernobyl nuclear power plant disaster in the Soviet Union, the most telling have come from its leader, Mr. Gorbachov. He said what happened showed again what an abyss will open if nuclear war befalls mankind.

The world has indeed been given a most graphic ___26___. There was a relatively small explosion in the Chernobyl ___27___ station and it took place in a reaction encased ___28___ steel and concrete. Even so, at least seven are ___29___ hundreds are in hospital and thousands have had to ___30___ evacuated.

The effects and consequences are inifinitesimal compared with ___31___ would have happened had a nuclear weapon been exploded ___32___ ground level. The magnitude of the danger is almost ___33___ beyond human comprehension.

As the Royal Society of New ___34___ observed last year we all know the difference between ___35___ and 100 deg. but what do we make of a ___36___ of 15 million degrees generated by a "small" atomic ___37___? How can we really appreciate the destructive capacity of ___38___ single modern nuclear bomb several times more powerful than ___39___ dropped on Hiroshima?

The stockpile of nuclear weapons distributed ___40___ the world contains the explosive power of more than ___41___ million Hiroshima bombs. They pose an unprecedented threat to ___42___ yet somehow the world lives with it. Perhaps the ___43___ of destruction is so awesome we cannot even begin ___44___ comprehend it.

An explosion in a nuclear power station ___45___ something we can understand, and worry about. At Chernobyl ___46___ a little of the nuclear genie escaped from the ___47___ by accident. When we see the fearful results we ___48___ begin to understand what would have happened if it ___49___ been nuclear weapons that exploded.

Hopefully the lesson from ___50___ will not be lost on all the peoples of ___51___ world and they will press even harder for nuclear ___52___. The arms race is still being pursued to the ___53___ of madness. Now is the time to end it.

242.

Write your answers next to the question numbers (below) which correspond to the numbered gaps in the passage.

26.

27.

28.

29.

30.

31.

32.

33.

34.

35.

36.

37.

38.

39.

40.

41.

42.

43.

44.

45.

46.

47.

48.

49.

50.

51.

52.

53.

SCIENCE COMPREHENSION

PASSAGE 4.

In different environments different kinds of animal communities exist, and in each of these are interrelationships based upon the food habits of the various species.

In a pond or small lake, for example, certain ___54___ animals (protozoa) feed upon bacteria, microscopic plants (algae), and ___55___, usually smaller, protozoa. The protozoa, in turn are fed ___56___ by small many-celled animals, such as rotifers and ___57___ crustacean relatives of the crayfish and the crab. These ___58___ turn, may be fed upon by larger crustaceans, by ___59___-living insects, and by small fishes. And these, in ___60___, may be fed upon by still larger fishes. Fish-___61___ birds may also be involved in this group of ___62___ of the hunter and the hunted. On land the ___63___ herbivores (mouse, squirrel, rabbit) are fed upon by the ___64___ carnivores (owl, hawk, fox), while the larger herbivores (cattle, ___65___, deer) are preyed upon by the larger carnivores (wolf, ___66___, lion).

Such a pattern of interrelationships among animals is ___67___ a food chain. Each food chain, whether composed of ___68___ links or only a few, is based upon the ___69___ that the predator is usually larger than its prey, ___70___ that the lowest link is made up of herbivorous ___71___.

It is obvious that the smallest animals in a ___72___ food chain relationship must be the most numerous, and ___73___ the number of animals in each successive link of ___74___ chain must be fewer and fewer. A small pond ___75___ support only a few fishes, and it might require ___76___ such ponds to support a single fish-eating bird. ___77___ in keeping with these relationships are the generally more ___78___ reproductive and growth rates of the smaller species.

Write your answers next to the question numbers (below) which correspond to the numbered gaps in the passage.

54.

55.

56.

57.

58.

59.

60.

61.

62.

63.

64.

65.

66.

67.

68.

69.

70.

71.

72.

73.

74.

75.

76.

77.

78.

PASSAGE 5.

Those of us who use hard coal as fuel in our furnaces can learn much while stoking the fire.

First, we may notice that when coal is put into the furnace and the bottom draught is opened, a brisk yellow blaze appears above the coals. Sparks leap upward and vanish. The luminosity of the blaze is mostly due to glowing particles of coal.

After the fire has a good start it is ___79___ to close the bottom draught and to open a ___80___ furnace door. In this way the fire is kept ___81___ burning too fast. In a little while a blue ___82___ can be seen, not burning steadily in any one ___83___ but disappearing and then reappearing with a gentle pop ___84___ puff. There is no yellow flame now, and there ___85___ no sparks flying upward.

What is the explanation of ___86___ simple facts? When carbon, or coal, burns in a ___87___ draught, the following reaction takes place:

$$C + O_2 \rightarrow \text{\_\_\_88\_\_\_}$$

That is, one atom of carbon unites with two ___89___ of oxygen to form one molecule of gas ___90___ as carbon dioxide. This gas is not inflammable, has ___91___ smell, and is harmless under most circumstances. It passes ___92___ the chimney and is lost in the air.

When ___93___ bottom draught is closed, very little air comes up ___94___ the layers of coals. Under these conditions, instead of ___95___ atoms of oxygen for each atom of carbon, the ___96___ requires only one. The process is:

$$C + O \text{\_\_\_97\_\_\_} CO$$

Since there are two oxygen atoms to a ___98___, the chemist would rather write this:

$$2C + O_2 \text{\_\_\_99\_\_\_} 2CO$$

With the draught closed, then one atom of ___100___ unites with one atom of oxygen to form one ___101___ of carbon monoxide, or CO. Carbon monoxide is a ___102___ gas without smell or taste, invisible, and therefore especially ___103___. It would be foolhardy to close the bottom draught ___104___ for the fact that carbon monoxide is inflammable. On ___105___ of the coals the carbon monoxide burns to the ___106___ harmless carbon dioxide. It burns on top of the ___107___ only because plenty of oxygen coming through the open ___108___ can reach it there.

PASSAGE 5 (Cont.)

The flame flickers because, after ___109___ for a moment, all the CO is gone from ___110___ one spot, and the flame goes out. Then the ___111___ draught moving upward brings some more CO; and when ___112___ open air is reached, the blue flame appears once ___113___.

Write your answers next to the question numbers (below) which correspond to the numbered gaps in the passage.

79.

80.

81.

82.

83.

84.

85.

86.

87.

88.

89.

90.

91.

92.

93.

94.

95.

96.

97.

98.

99.

100.

101.

102.

103.

104.

105.

106.

107.

108.

109.

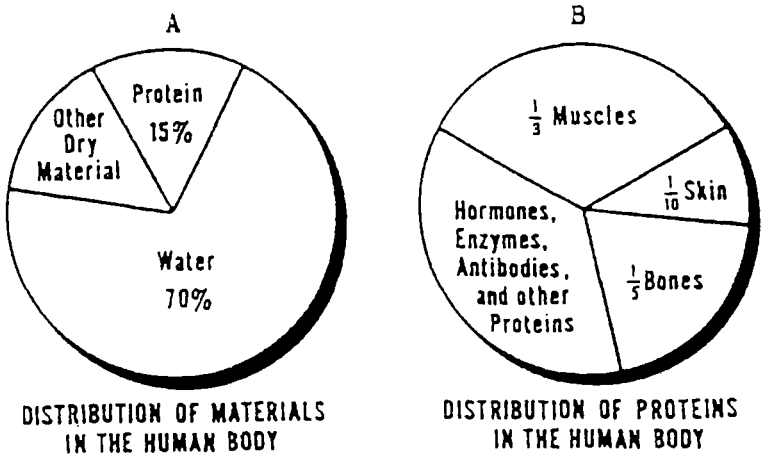
110.

111.

112.

113.

PASSAGE 6.



DISTRIBUTION OF MATERIALS IN THE HUMAN BODY

DISTRIBUTION OF PROTEINS IN THE HUMAN BODY

NOTE: Graph B is an "expanded" view of the protein section in Graph A.

114. The human body, according to the data furnished by the two graphs, is composed mainly of what substances?

115. A person weighing 80kg would, according to these graphs, be composed of water weighing how many kilograms?

116. In the human body, the distribution of material other than water and protein is equal to (in its simplest form)

117. How many degrees of the circle should be used to represent the distribution of protein? (NOTE: there are 360^{o} in a circle.)

118. The ratio of the distribution of proteins in muscle to the distribution of proteins in skin is (in its simplest form)

SOCIAL STUDIES COMPREHENSION

PASSAGE 7.

The natural environment, commonly termed the land, is the agriculturalist's basic resource. The agriculturalist has the choice of using this resource ___119___ two fundamental ways: he may regard the land as ___120___ fund type resource to be exploited before moving to ___121___ area; or, as a flow type resource to be ___122___ in such a way as to maintain its productivity ___123___. The inherent qualities of the land, the pressure of ___124___ on the land, and the degree of technological skill ___125___ cultural background of the agriculturalist are some of the ___126___ factors affecting his response to this choice. To produce ___127___, in each case, the agriculturalist applies labour and capital ___128___ the land in varying proportions and intensities.

The products ___129___ from his plants and animals provide the agriculturalist with ___130___ means of livelihood. These products may be consumed directly, ___131___ the agriculturalist and his family, or indirectly after some ___132___ of processing to give the original product form utility, ___133___ make the good into a more acceptable form for ___134___. A large volume of the world's agricultural products are ___135___ consumed by those agriculturalists who produce them but are ___136___ or sold to non-agriculturalists. These products have to ___137___ transported, and perhaps stored and processed before consumption. Transport ___138___ place utility, that is, the product is made more ___139___ by transfer to a location where it is demanded, ___140___ storage may add time utility by providing the good ___141___ it is wanted. Agricultural products which are processed, transported, ___142___ stored, gain value well beyond that which the agriculturalist ___143___ for them because the costs of utility are added ___144___ the original product.

250.

Write your answers next to the question numbers (below) which correspond to the numbered gaps in the passage.

119.

120.

121.

122.

123.

124.

125.

126.

127.

128.

129.

130.

131.

132.

133.

134.

135.

136.

137.

138.

139.

140.

141.

142.

143.

144.

PASSAGE 8

To discuss criminology in the Third World in isolation from the fundamental reality of imperialism would be fruitless and totally inadequate and misleading. My argument is that crime is not an autonomous ___145___ which applies to certain kinds of behaviour in all ___146___ and cannot therefore be studied in itself. Rather it is socially determined, essentially by the powerful groups in society ___147___ their own purposes; therefore anyone wishing to consider seriously ___148___ concept of crime must analyse the following: (a) what ___149___ have power in society; (b) the purpose for which ___150___ power is exercised in relation to the designation of ___151___ as criminal. Such an analysis will indicate that by ___152___ large a minority with power in society determine which ___153___ shall be considered criminal.

In the case of colonial ___154___, the power lay with the colonial administrators and they ___155___ it partly through the legal system, primarily against the ___156___ powerless indigenous population. There can be no serious argument ___157___ the view that the fundamental purpose of the colonial ___158___ system was to repress the indigenous population in order ___159___ allow the colonial link to be maintained, for whatever ___160___, strategic (or combination of) reasons.

And I have recently ___161___ elsewhere that the study of crime in colonial Africa ___162___ part of a wider intellectual endeavour which was intended ___163___ assist the colonial administration in maintaining its dominance. Numerous ___164___ have argued recently that criminology serves a similar purpose ___165___ capitalist countries, that is, it functions most importantly to ___166___ in maintaining the status quo, by providing the justification ___167___ repression of the working class both physically and ideologically.

252.

Write your answers next to the question numbers (below) which correspond to the numbered gaps in the passage.

145.

146.

147.

148.

149.

150.

151.

152.

153.

154.

155.

156.

157.

158.

159.

160.

161.

162.

163.

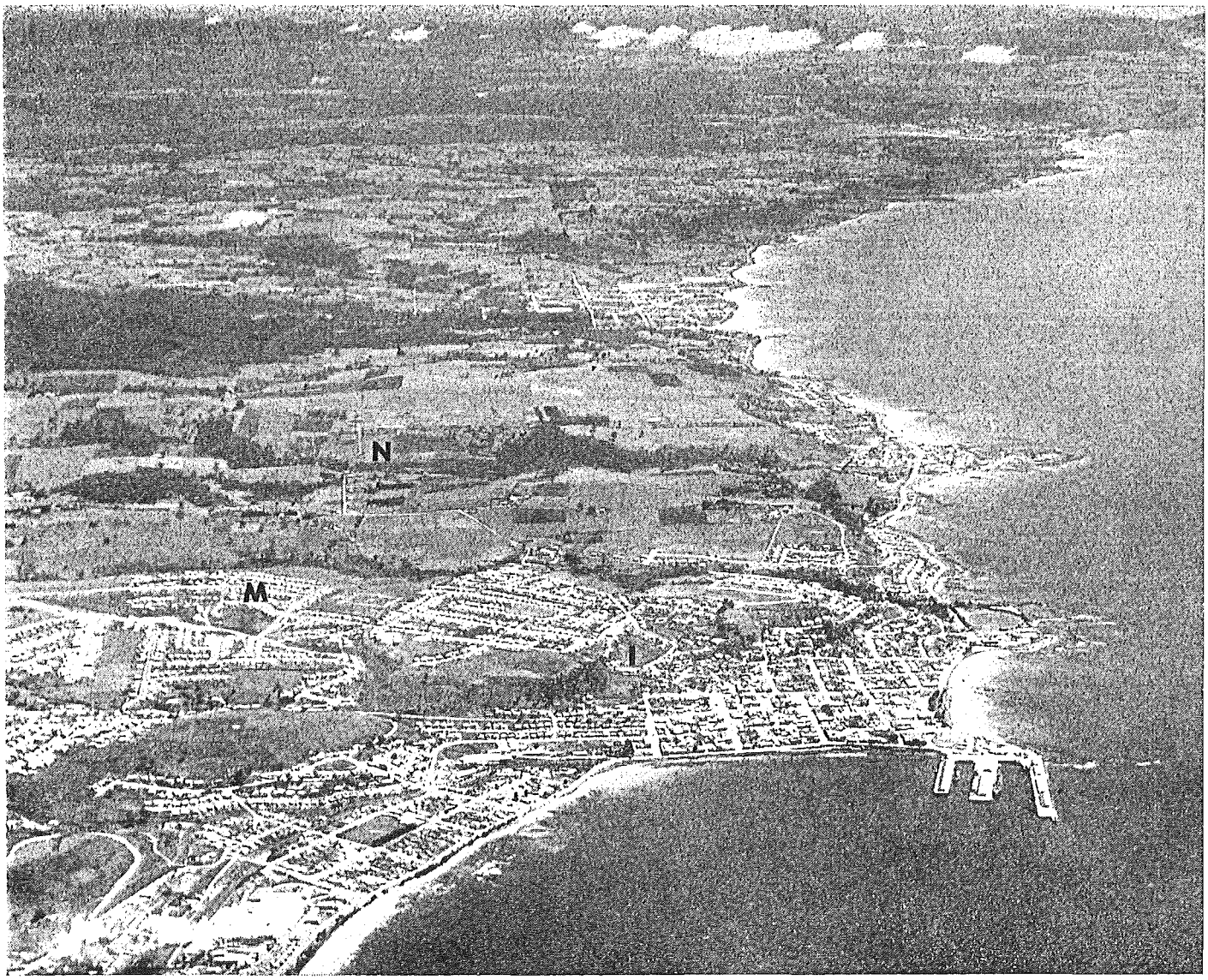
164.

165.

166.

167.

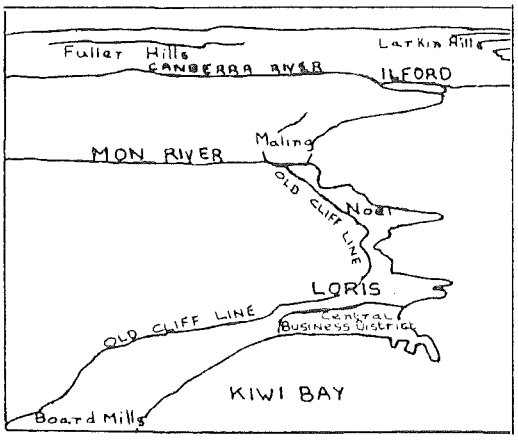
PASSAGE 9.



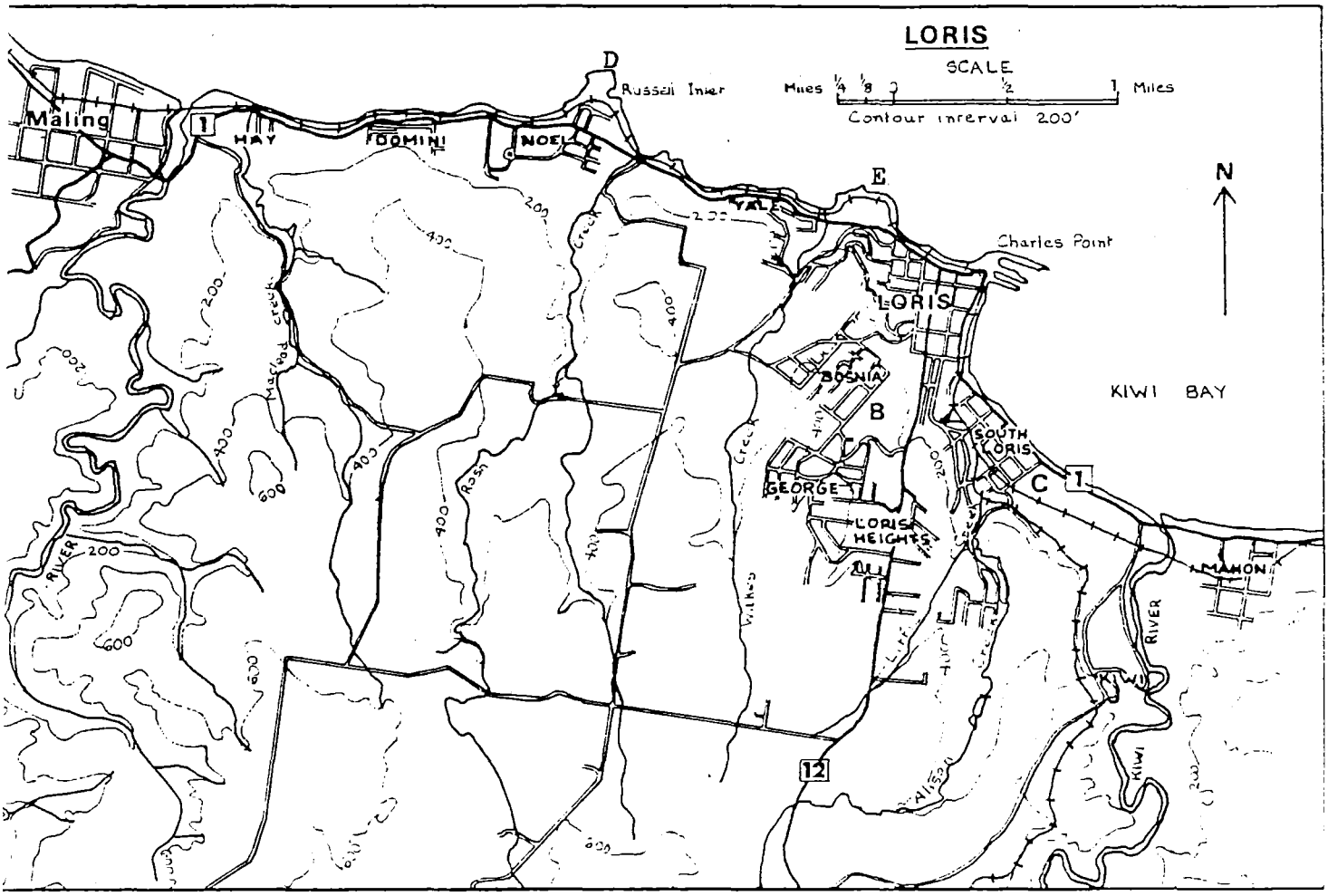
Loris and the North-west Coast

The photograph looks westward and shows a considerable section of the northern coastland west of Loris. The soils on the plateau are used for dairy farming and the growing of cash crops, notably potatoes. Although the plateau slopes more or less gently to the sea, an old abandoned sea cliff occurs some distance behind the present shore along the greater part of the coast. As well as being an important service centre, Loris has developed as an industrial town. A major industry concerned with the manufacture and processing of wood pulp is prominent in the bottom left hand corner of the photograph. Since the photograph was taken, the port has been extended by the construction of new breakwaters protecting the harbour area from the north-east.

Sketch of the Area Shown in the Photograph



Map of Loris



LEGEND

0 to 25,000	SLADE
0 to 5,000	Dorset
s than 1,500	PEARCE

Railways	
Route No.	3
Roads	
Creeks	
Dams	
Contour lines	200

NOTES

tours are lines on a map joining places of the same height above sea level.
base (0 metres) is sea level. The constant difference in height between one
and the next is known as the contour interval.

168. If a man left the township of Noel and drove eastward along Route 1 for two miles he would be in which town?

169. Locate on the map the site marked "I" in the photograph. The site marked "I" is most probably what?

170. The settlement pattern of the inland area between Wilkes Creek and the Mon River is best described as consisting of

171. A train follows the Kiwi River valley after leaving South Loris. While in the area shown in the map and photograph, it would most probably travel across what type of physical terrain?

172. From the information on the map what can we conclude about the population of Bosnia in relation to the population of Mahon?

IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

FORM A

MULTIPLE CHOICE

MATHEMATICS

TEST

INSTRUCTIONS

Each of the questions or incomplete statements is followed by suggested answers. You are to choose the BEST answer (A,B,C,D, or E) and show your choice by putting a CIRCLE around the appropriate letter on the separate ANSWER SHEET.

EXAMPLE QUESTION:

> The sum of 6 + 8 + 10 is
>
> A 12 B 18 C 24 D 30 E 34
>
> The sum of 6, 8, and 10 is 24, so the correct answer is C. It is CIRCLED on the ANSWER SHEET like this:
>
> A B (C) D E

Mark only ONE answer for each question.

There are 25 questions to be answered.

Do NOT spend too much time on any one question.

You will have 25 minutes to complete the test.

Do NOT write in the question booklets. (You may use a piece of scrap paper for working out if you wish.)

DO NOT START UNTIL YOU ARE TOLD

1. The map scale reads "1 cm = 3 km" Linda finds that two points on the map are 15cm apart. How many kilometres is that?

A 5

B 9

C 15

D 30

E 45

2. Which of the following represents the number which is 7 greater that 11?

A $7 > 11$

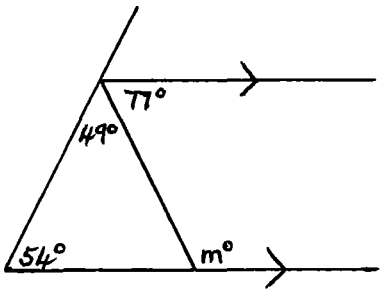
B $11 > 7$

C 11×7

D $11 + 7$

E $11 - 7$

3. The value of angle m in the following diagram is



A 49°

B 54°

C 77°

D 103°

E 131°

4. If $t + 1 = \frac{4t - 5}{2}$, $t = ?$

A $-1\frac{1}{2}$

B 1

C 3

D $3\frac{1}{2}$

E 6

5. If N is equal to $\frac{4}{5}$ of the average of the number 7, 9, and 14, then N = ?

A 8

B 10

C 12

D 12.5

E 24

6. A family saves 5% of its monthly income. If the monthly income is increased from $600 to $650, by how much are the monthly savings increased?

A $5.00

B $2.50

C $1.75

D $1.50

E $1.00

7. One economical method of multiplying a large number such as 6,323 by 25 requires two steps. The first step is to multiply 6,323 by 100. What is the second step?

A Divide the result of the first step by 5.

B Divide the result by 4

C Divide the result by 3

D Subtract 7500 from the result

E Subtract 75 from the result

8. A man owes a debt of $110. If he makes weekly payments of $5 each, how many payments must he make before the remainder of the debt will be less than the amount he has already paid?

A 10

B 11

C 12

D 13

E 15

9. The decimal numeral for $7 + \frac{2}{100} + \frac{3}{1000}$ is

A 7.023

B 70.23

C 702.3

D 7.23

E 72.3

10. $\frac{9}{12} - \frac{6}{8} = ?$

A 0

B $\frac{1}{8}$

C $\frac{9}{16}$

D $\frac{3}{4}$

E 1

11. How many cubes with edges 2cm long can be made from a block 6cm long, 4cm wide and 2cm high?

A 48

B 24

C 12

D 10

E 6

12. A certain company had 2700 employees, of whom $\frac{1}{3}$ were hand operators. Upon hiring 200 machine operators, the company dismissed 400 hand operators. After this change, what fraction of the company's employees were hand operators?

A $\frac{5}{27}$

B $\frac{3}{16}$

C $\frac{1}{5}$

D $\frac{2}{9}$

E $\frac{2}{7}$

13. The table below shows the number of eggs laid each week by a farmer's ten hens:

EGGS	HENS
1	2
3	2
4	1
6	2
7	3

What percentage of hens laid more than four (4) eggs in the week?

A 60%

B 50%

C 40%

D 30%

E 17%

14. A family left on a trip at 8:15 a.m. They arrived at their destination at 2:38 p.m. How long did it take to make the trip?

A 5 hours, 23 minutes

B 6 hours, 37 minutes

C 6 hours, 53 minutes

D 10 hours, 53 minutes

E None of the above.

15. Simplify $2\frac{1}{4} + \frac{5}{6} + 1\frac{5}{8}$

A $3\frac{25}{192}$

B $1\frac{17}{24}$

C $3\frac{11}{24}$

D $4\frac{17}{24}$

E $3\frac{11}{18}$

16. The number of metres (N) that a free object will fall in T seconds is indicated by the formula $N = 16\ T^2$. If a lead weight is dropped from the top of a 576 metre tower, how long will it take to fall to the ground?

A $3\sqrt{2}$ seconds

B $\sqrt{6}$ seconds

C 6 seconds

D 36 seconds

E None of the above.

17. The first four values in a series of numbers written according to a set scheme are 65, 50, 37, 26, 17. On the basis of the scheme used, what is the next value of this series?

A 4

B 6

C 8

D 10

E 12

Problems 18 and 19 are based on the following table.

Five brands of tyres were tested for durability under actual road conditions. The mileage record for ten tyres of each brand is indicated in the table below:

Mileage	Number of Tyres Giving Indicated Mileage				
	Brand				
	V	W	X	Y	Z
24,000 - 25,999 km		1	2		
22,000 - 23,999 km		2	6	2	3
20,000 - 21,999 km	3	4	2	3	4
18,000 - 19,999 km	4	2		3	3
16,000 - 17,999 km	3	1		2	
Total No. of Tyres	10	10	10	10	10

18. On the basis of this test which brand should be recommended?

A W

B X

C Y

D Z

E No brand was superior to the others.

19. For which brand of tyres did durability vary most markedly from one tyre to another?

A V

B W

C Y

D Z

E They all varied to about the same degree.

20. A portable television set may be purchased for $129.50 cash, or $50 deposit and $5 a week for 22 weeks. What is the additional cost which results from purchase on the installment plan?

A $21.50

B $31.50

C $41.50

D $50.50

E None of the above.

21. Suppose that the first astronaut to land on Mars finds that the Martians have three fingers on one hand and two on the other and that they have an additive numeration system.

If they wrote ✡✡✡∘ for sixteen

Ø✡∘ for fifty-one

Ø∘∘∘ for twenty-eight,

what does their numeral Ø✡∘ stand for?

A 13

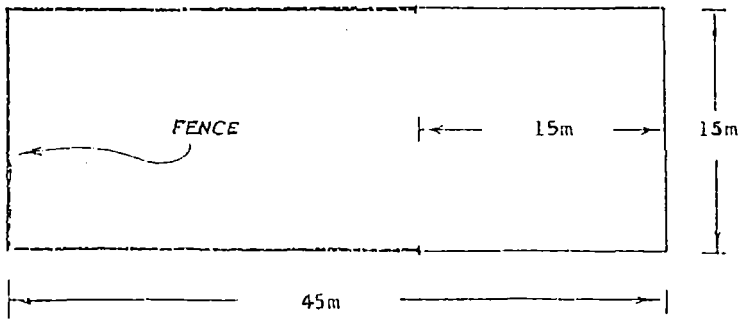
B 29

C 31

D 301

E 2551

22.



How many sheets of asbestos, each 1½ metres wide are needed to fill in the gap in the fence around the block of land shown in the plan above?

A 10

B 30

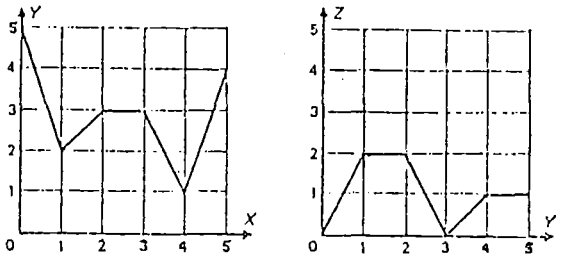
C 50

D 70

E 112½

Questions 23 and 24 refer to the following information.

The two graphs below show the relationship between X and Y, and between Y and Z.



23. What is the value of Z when X = 2?

A 0

B 1

C 2

D 3

E 4

24. When X increases from 3 to 4, what is the change in Z?

A an increase of 2

B an increase of 1

C a decrease of 2

D a decrease of 1

E there is no change

25. How many letters are there in SUT?

i.e. find n(SUT)

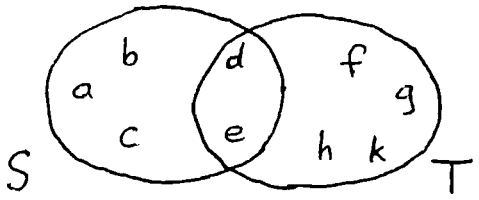
A 2

B 3

C 4

D 5

E 9



IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

NAME(please print):_____ APPENDIX G _____

AGE(in years):_____ SEX(circle one): M F

FORM CLASS:_____ SCHOOL:_____

FORM A

OPEN - ENDED MATHEMATICS TEST

INSTRUCTIONS

Write your answer to each question or incomplete statement in the QUESTION BOOKLET in the space provided. It will be to your advantage to show your working where possible.

EXAMPLE QUESTION:

The sum of 6 + 8 + 10 is

$$\begin{array}{r} 6 \\ 8 \\ +10 \\ \hline 24 \end{array}$$

The sum of 6, 8 and 10 is equal to 24, thus the correct answer is 24. This is how you should answer all the questions. REMEMBER to do all your working in the space provided.

There are 25 questions in total.

Do NOT spend too much time on any one question.

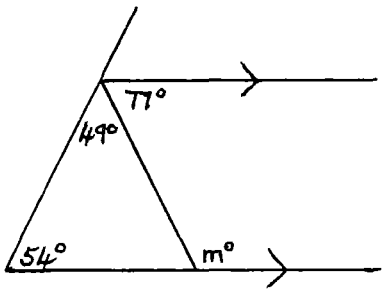
You will have 25 minutes to complete the test.

DO NOT START UNTIL YOU ARE TOLD

1. The map scale reads "1 cm = 3 km" Linda finds that two points on the map are 15cm apart. How many kilometres is that?

2. **What is the number which is 7 greater than 11 ?**

3. The value of angle m in the following diagram is



4. If $t + 1 = \frac{4t - 5}{2}$, $t = ?$

5. If N is equal to $\frac{4}{5}$ of the average of the numbers 7, 9, and 14, then N = ?

6. A family saves 5% of its monthly income. If the monthly income is increased from $600 to $650, by how much are the monthly savings increased?

7. One economical method of multiplying a large number such as 6,323 by 25 requires two steps. The first step is to multiply 6,323 by 100. What is the second step?

8. A man owes a debt of $110. If he makes weekly payments of $5 each, how many payments must he make before the remainder of the debt will be less than the amount he has already paid?

9. The decimal numeral for $7 + \frac{2}{100} + \frac{3}{1000}$ is

10. $\frac{9}{12} - \frac{6}{8} = ?$

11. How many cubes with edges 2cm long can be made from a block 6cm long, 4cm wide and 2cm high?

12. A certain company had 2700 employees, of whom $\frac{1}{3}$ were hand operators. Upon hiring 200 machine operators, the company dismissed 400 hand operators. After this change, what fraction of the company's employees were hand operators?

13. The table below shows the number of eggs laid each week by a farmer's ten hens:

EGGS	HENS
1	2
3	2
4	1
6	2
7	3

What percentage of hens laid more than <u>four</u> (4) eggs in the week?

14. A family left on a trip at 8:15 a.m. They arrived at their destination at 2:38 p.m. How long did it take to make the trip?

15. Simplify $2\frac{1}{4} + \frac{5}{6} + 1\frac{5}{8}$

16. The number of metres (N) that a free object will fall in T seconds is indicated by the formula $N = 16\ T^2$. If a lead weight is dropped from the top of a 576 metre tower, how long will it take to fall to the ground?

17. The first four values in a series of numbers written according to a set scheme are 65, 50, 37, 26, 17. On the basis of the scheme used, what is the next value of this series?

Problems 18 and 19 are based on the following table.

Five brands of tyres were tested for durability under actual road conditions. The mileage record for ten tyres of each brand is indicated in the table below:

Mileage	Number of Tyres Giving Indicated Mileage				
	Brand				
	V	W	X	Y	Z
24,000 - 25,999 km		1	2		
22,000 - 23,999 km		2	6	2	3
20,000 - 21,999 km	3	4	2	3	4
18,000 - 19,999 km	4	2		3	3
16,000 - 17,999 km	3	1		2	
Total No. of Tyres	10	10	10	10	10

18. On the basis of this test which brand should be recommended?

19. For which brand of tyres did durability vary most markedly from one tyre to another?

20. A portable television set may be purchased for $129.50 cash, or $50 deposit and $5 a week for 22 weeks. What is the additional cost which results from purchase on the installment plan?

21. Suppose that the first astronaut to land on Mars finds that the Martians have three fingers on one hand and two on the other and that they have an additive numeration system.

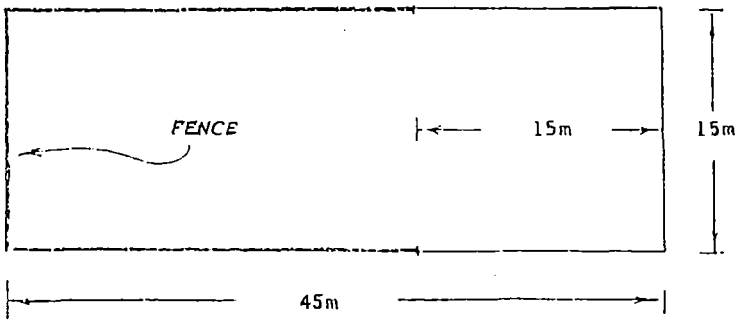
If they wrote ✡✡✡○ for sixteen

⌀⌀○ for fifty-one

⌀○○○ for twenty-eight,

what does their numeral ⌀✡○ stand for?

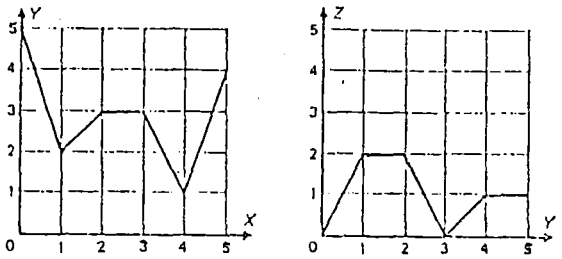
22.



How many sheets of asbestos, each 1½ metres wide are needed to fill in the gap in the fence around the block of land shown in the plan above?

Questions 23 and 24 refer to the following information.

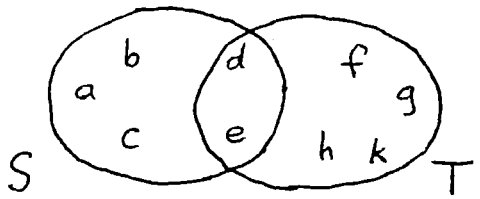
The two graphs below show the relationship between X and Y, and between Y and Z.



23. What is the value of Z when X = 2?

24. When X increases from 3 to 4, what is the change in Z?

25. How many letters are there in SUT?
i.e. find n(SUT)



IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

APPENDIX H

ESSAY TEST

This test attempts to assess your ability to write an essay that reflects the four following characteristics:

* interest and liveliness
* correct grammar, spelling and punctuation
* a clear and fluent writing style
* appropriate organization.

INSTRUCTIONS

Choose ONE of the following six topics and write three or four paragraphs on it (about 200 words).

Write the essay on a SEPARATE SHEET. REMEMBER to record your NAME and TOPIC NUMBER at the top of the answer sheet.

TOPICS:

If you choose topics 1 or 2, try to persuade the reader that your opinion is right.

1. Young people are not given enough freedom.

2. Should students wear school uniform?

OR use topics 3 or 4 as the beginning sentence of an essay.

3. "There, in the distance we could just see...."

4. "They saw the gap, and decided it was now or never...."

OR for topics 5 or 6 describe how you felt.

5. What should I do now?

6. An unpleasant experience.

You will have 25 minutes to complete your essay.

Please do NOT write on or mark this question sheet.

DO NOT START UNTIL YOU ARE TOLD

APPENDIX H-1

ESSAY MARKING SCHEDULE

1. MECHANICS: (5 marks)

(i) Length
- 1 mark if essay is 180 words or more;
- ½ mark if essay is between 100-179 words; or
- 0 mark if essay is less than 100 words.

(ii) Opening
- ½ mark for a lively and interesting first sentence; or
- 0 mark for a dull and boring first sentence.

(iii) Conclusion
- ½ mark for a neat and tidy summary in final sentence(s); or
- 0 mark for a dull and inconclusive final sentence(s).

(iv) Paragraphing
- 1 mark for logical and appropriate use of paragraphs; or
- 0 mark for essay without paragraphs or poor use of.

(v) Grammar
- 2 marks if essay has only 0-4 errors;
- 1 mark if essay has between 5-9 errors; or
- 0 mark if essay has 10 or more errors.

(N.B. Grammatical errors to be assessed on the first 150 words only).

2. CONTENT: (5 marks)

Definition: The colourfulness and depth of vocabulary regardless of its grammatical correctness.

5 marks - excellent; top class.
4 marks - very good; but not quite top class.
3 marks - average; sound but rather lifeless.
2 marks - poor; but without gross deficiencies.
1 mark - very poor; minimum effort shown.
0 mark - hopeless; either nothing or only a few lines written.

(N.B. ½ marks allowed)

3. STYLE: (5 marks)

Definition: The use of appropriate language, as well as displaying originality in both ideas and expression.

5 marks - excellent; top class.
4 marks - very good; but not quite top class.
3 marks - average; but rather repetitive and/or common.
2 marks - poor; but without gross deficiencies.
1 mark - very poor; minimum effort shown.
0 mark - hopeless; either nothing or only a few lines written.

(N.B. ½ marks allowed)

4. ORGANISATION: (5 marks)

Definition: Logical and effective use of facts and ideas. Does the argument or storyline flow in a logical manner?

5 marks - excellent; top class.
4 marks - very good; but not quite top class.
3 marks - average; sound with no glaring mistakes.
2 marks - poor; but still showing some logic.
1 mark - very poor; minimum effort shown.
0 mark - hopeless; either nothing or only a few lines written.

(N.B. ½ marks allowed)

APPENDIX I

GENERAL SCHOLASTIC APTITUDE TEST

This test attempts to measure how well you think in relation to common school-related tasks. The test consists of different kinds of word and number problems.

INSTRUCTIONS

There are four sections in this test. For each question or incomplete statement you are to choose the BEST answer (A,B,C,D or E) and show your choice by CIRCLING the appropriate letter on the separate ANSWER SHEET. Select only ONE answer for each question.

EXAMPLE QUESTION:

How many months are there in a year?

A 9
B 10
C 11
D 12
E None of the above

There are 12 months in a year, so answer D is the correct choice. Therefore letter D has been circled on the answer sheet. This is how you should respond to each question.

If you want to CHANGE an answer, simply cross out the error and clearly indicate your new choice.

At the beginning of each new section, there is a practice example for you to study.

There are 25 questions in total.

Do NOT spend too much time on any one question.

You will have 25 minutes to complete the test.

Please do NOT write in or mark this question booklet.

DO NOT START UNTIL YOU ARE TOLD

SECTION ONE

This section is concerned with questions of word knowledge. For each question, choose the answer which has the SAME meaning as the UNDERLINED word.

EXAMPLE S1:

Quiet means

A blue
B still
C tense
D watery
E exact

The word "still" means the same as the underlined word quiet, therefore option B is the correct answer. The letter B has been CIRCLED on the answer sheet.

1. Oblivious
   A painless
   B static
   C eternal
   D dead
   E unmindful

2. Obscurity
   A poverty
   B unknown
   C disgrace
   D decay
   E passion

3. Premature
   A unfortunate
   B tardy
   C prejudiced
   D final
   E early

4. Trivial

A good-natured
B heated
C unimportant
D bitter
E one-sided

5. Erroneous

A strange
B outrageous
C a consequence
D a retaliation
E a repercussion

6. Divulge

A discover
B shield
C gossip
D reveal
E converse

For questions 7 to 9, choose the ONE answer which gives the OPPOSITE meaning to the underlined word.

7. Obliterate

A construct
B alter
C expose
D erase
E object

8. Irrational

A mistaken
B stubborn
C logical
D unreasonable
E sensitive

9. Placid

A unfriendly

B bashful

C restless

D amiable

E lazy

SECTION TWO

These questions show how well you understand the relationship between words.

EXAMPLE S2:

LIGHT is to DARK as PLEASURE is to

A picnic
B day
C pain
D night
E boat

The correct answer is C because pain is the opposite of pleasure just as dark is the opposite of light. Therefore the letter C has been circled on the answer sheet.

10. MIDNIGHT is to 4 A.M. as 10 A.M. is to

A 6 a.m.
B 9 a.m.
C NOON
D 2 p.m.
E 6 p.m.

11. ORDER is to OBEY as DEMAND is to

A insist
B comply
C deny
D object
E submit

12. BUY is to PERMANENT as RENT is to

A convenient
B certain
C difficult
D flat
E temporary

13. PUTTER is to CUE as GOLF is to
    A actors
    B billiards
    C badminton
    D play
    E archery

14. LEGAL is to LAWYER as REGAL is to
    A minister
    B regalia
    C king
    D royal
    E earl

15. FACTORY WORKER is to WAGE as MANAGER is to
    A profession
    B overtime
    C executive
    D salary
    E money

16. Which of the following means the SAME as "Don't put all your eggs in one basket"?
    A It is safer not to risk all on one venture.
    B It is easier to carry all your eggs in two baskets than in one.
    C Division of responsibility brings poor results.
    D It is better to be content with what you have than to lose it gambling for more.
    E None of the above.

SECTION THREE

The questions in this section measure how well you can follow a series of numbers. Respond by choosing the ONE answer which BEST fits the pattern of the number series.

EXAMPLE S3:

1 3 5 7 9 11

A 12

B 13

C 14

D 15

D None of the above

The correct answer is B. The number pattern consists of successive odd numbers therefore the next member of the series will be 13. The letter B has been circled on the answer sheet.

17. 10 11 8 9 6 7 4 5 2

A 2

B 3

C 4

D 5

E 6

18. 1 11 20 28 35 41 46

A 49

B 50

C 52

D 54

E 57

19. 0 1 0 0 2 0 0 0 4 0 0 0

A 0

B 2

C 4

D 8

E 16

20. 1 3 5 8 11 15 19 24 29

A 32

B 33

C 34

D 35

E 36

21. 1 3 7 15 15 7

A 7

B 6

C 5

D 4

E 3

SECTION FOUR

This section is concerned with questions of verbal reasoning. Respond by choosing the ONE option which you consider to be the BEST answer.

EXAMPLE S4:

John runs faster than Fred. Fred runs faster than Ian. Tony and Sid are slower than Ian. Who runs the fastest?

A John

B Fred

C Ian

D Tony

E Sid

As John runs faster than both Fred and Ian; and Tony and Sid are slower than Fred or Ian then John is the fastest. Therefore answer A is correct. The letter A has been circled on the answer sheet.

22. Which of the following is FALSE?

Of a company's fleet of cars, few are automatic, but all have four wheel brakes. Therefore,

A some cars have four wheel brakes and are automatic.

B some cars have four wheel brakes but are not automatic.

C all cars have four wheel brakes or are automatic but not both.

D of those cars which have four wheel brakes, few are also automatic.

E those cars which are automatic also have four wheel brakes.

23. In a bag are 50 kilograms of oranges and 20 kilograms of lemons and in all, 30 kilograms of fruit are not good. What is the greatest possible weight of good oranges in the bag?

A 10kg

B 20kg

C 30kg

D 40kg

E 50kg

24. Peter has a half-holiday on Wednesday and Saturday afternoons, and a whole day holiday on Sunday. I am at work all day, except on Monday, Wednesday, Friday and Sunday. I want to take Peter shopping to buy a new suit. Which afternoon could we go together?

A Monday

B Tuesday

C Wednesday

D Thursday

E Friday

25. My grocer promised to deliver my groceries at about 10 o'clock unless he was still at the market, in which case he would come at about eleven. It is five past ten and there is a knock at the back door. Of which of the following can I be certain?

A The grocer has left the market.

B If it is not the grocer he has broken his promise.

C The grocer was detained for a while at the market.

D The grocer may not be coming for another hour.

E The grocer has brought my groceries.

IF YOU FINISH EARLY, GO BACK AND CHECK YOUR ANSWERS

APPENDIX J

LIST OF ITEM SOURCE MATERIALS

1. Academic Promise Test (APT), Forms A and B (Verbal and Numerical Sub-Tests). The Psychological Corporation, New York, 1965.

2. Commonwealth Secondary Scholarships Examination (CSSE), Comprehension and Interpretation (Humanities). ACER, Hawthorn, Victoria, 1969.

3. Extension Studies English Test Booklet, University of the South Pacific, Extension Services, 1981. (Unpublished).

4. General Aptitude and Mathematics Reference Tests, Gilmore, A.M. Unpublished Ph.D. Thesis (Education), University of Otago, New Zealand, 1979.

5. Iowa Tests of Education Development (ITED). Forms X-3s and Y-3s; Form Y-4 (Grades 9-12). Science Research Associates, Chicago, Illinois, 1959; 1963.

6. Kaiapoi High School, Form 4 End of Year Mathematics Examination, Kaiapoi, New Zealand, 1977. (Unpublished)

7. Primary Mental Abilities (PMA). Grades 9-12, Science Research Associates, Chicago, Illinois, 1962.

8. Proficiency in English Measure (PEM). University of the South Pacific, Unpublished and Undated.

9. Scholastic Aptitude Test (SAT). Form BSAI, (Verbal and Mathematics Sub-Tests). College Entrance Examination Board, Educational Testing Service, Princeton, N.J. Undated.

10. School Certificate English Examination. Department of Education, Wellington, New Zealand, 1978.

11. Sequential Tests of Educational Progress (STEP). Grade 2B. Educational Testing Service, Princeton, N.J., 1962.

12. Test Of Scholastic Abilities (TOSCA). Secondary Grade, Form B. NZCER, Wellington, New Zealand, 1981.

APPENDIX K

ADMINISTRATION OF FORM 5 REFERENCE TESTS
INSTRUCTIONS FOR SUPERVISORS

1. Justification for the Testing:

   1.1 Research Purposes (Testing of a potential method of standardizing teachers' grades between classes, schools and subjects with a view to the introduction of full internal assessment at Forms 5 and 6).

   1.2 Return of Test Results to the School; and

   1.3 Good Practice for Own School Certificate and School-based Examination.

2. Testing Details:

   2.1 Tests in MULTIPLE CHOICE and OPEN-ENDED/CLOZE formats;

   2.2 The tests are already in a prearranged order; simply hand out in a random fashion;

   2.3 REMEMBER to ask pupils to check that all multiple-choice tests have a separate answer sheet;

   2.4 Tell pupils to FILL IN NAME, CLASS ETC. as specified;

   2.5 REPLACE any incomplete or spoiled papers with the EXACT same test format (MC or OE/C) and form (A or B);

   2.6 Read DUAL INSTRUCTIONS (MC and OE/C) for Vocabulary, Comprehension and Mathematics Tests. Single set of instructions for Essay and GSAT Tests;

2.7 REMIND pupils of TIME LIMIT for each test and write TIME CHECKS on blackboard;

2.8 DURING TEST check that pupils are answering respective item formats in correct manner;

2.9 Even if a pupil is NOT TAKING A PARTICULAR SUBJECT he/she should still attempt the test.

APPENDIX L

INTERCORRELATIONS OF THE REFERENCE TESTS

	Engl Vocab	Scie Vocab	SoSt Vocab	Engl Comp	Scie Comp	SoSt Comp	Maths	Essay	GSAT
Engl Vocab	-	.61	.67	.55	.59	.66	.49	.40	.61
Scie Vocab		-	.66	.49	.56	.58	.55	.30	.56
SoSt Vocab			-	.53	.55	.59	.49	.35	.59
Engl Comp				-	.60	.66	.43	.38	.60
Scie Comp					-	.73	.57	.29	.64
SoSt Comp						-	.55	.30	.63
Maths							-	.21	.64
Essay								-	.40
GSAT									-

All r's statistically significant at $p < 0.01$

REFERENCES

ADAMS, R.J. (1984) Sex Bias in ASAT? Hawthorn, Victoria; Australian Council for Educational Research.

AIKEN, JR., L.R. (1965) The Probability of Chance Success on Objective Test Items. Educational and Psychological Measurement, 25(1): 127-34.

ALDERSON, J.C. (1978) Critique of the Cloze Procedure and What it Supposedly Measures. IN Buros, O.K. (Ed.) Eighth Mental Measurements Yearbook, Highland Park, N.J.; Gryphon Press: 1171 - 4.

ANDERSON, T.H. (1974) Cloze measures as Indicies of Achievement Comprehension When Learning from Extended Prose. Journal of Educational Measurement, 11(2): 83-92.

ANDERSON, R.C. and FREEBODY, P. (1981) Vocabulary Knowledge. IN Guthrie, J.T. (Ed.) Comprehension and Teaching: Research Reviews. Newark, Delaware; International Reading Association.

ARCHER, J. (1984) The Implementation of an Education Innovation: The Introduction of School Based Assessment in Secondary Schools in Queensland. Unpublished M. Ed. Thesis, University of Queensland.

AUSTRALIAN SCHOLASTIC APTITUDE TEST (Undated) Test Specification. Hawthorn, Victoria; Unpublished Paper, Australian Council of Educational Research.

BAGNALL, M.F. and D'CRUZ, J.V. (1971) Moderating: Standardization by Consensus in Education, a Victorian Experiment in Teacher Professionalism. Australian Journal of Education, 15(1):104-17.

BALDAUF, JR., R.B. (1980) Why do Educational Measurement Texts Omit the Cloze Procedure? Educational and Psychological Measurement, 40(4):931-8.

BELL, R.C. (1973) Reliability, Item Analysis and Total Test Characteristics of ASAT-B. Hawthorne, Victoria; Unpublished paper, Australian Council for Educational Research.

BELL, R.C. (1977) A Psychometric Study of the ASAT (Series B). Nedlands, WA; University of Western Australia, Research Unit in University Education.

BERKELEY, G.F. and ALFORD, N.D. (1974) A Study of Developed Abilities. IN Dunn, S.S. (Ed.) Public Examinations: The Changing Scene. Adelaide, SA; Rigby:111-35.

BLACK, D.B. (1960) The Prediction of Freshman Success. Alberta Journal of Educational Research, 6(1):38-53.

BLOOM, B.S. (1965), (Ed.) Taxonomy of Educational Objectives: The Classification of Educational Goals. IN Handbook I: Cognitive Domain, N.Y.; Longmans.

BORMUTH, J.R. (1965) Comparisons Among Cloze Test Scoring Methods. Paper read at the Annual Convention of the California Educational Research Association, March 12.

BORMUTH, J.R. (1967) Comparable Cloze and Multiple-Choice Comprehension Test Scores. Journal of Reading. 10:291-9.

BORMUTH, J.R. (1968) Cloze Test Readability: Criterion Reference Scores. Journal of Educational Measurement, 5:189-96.

BORMUTH, J.R. (1969) Factor Validity of Cloze Tests as Measures of Reading Comprehension. Reading Research Quarterly, 4(3):358-65.

BORTNIK, R. and LOPARDO, G.S. (1973) An Instructional Application of the Cloze Procedure. Journal of Reading, 16(4):296-300.

BRAY, D.H. (1971) Examinations and Internal Assessment: A New Proposal. Post Primary Teachers' Association Journal, 18(5):37-40.

BRISTOW, S.D. (1971) Examinations and the Future. Post Primary Teachers' Association Journal, 18(6):32-4.

BROWN, J. (1966) Objective Tests: Their Construction and Analysis. London; Longmans.

CAPPER, P. (1987) Do or Die in the Senior School. Post Primary Teachers' Association Journal, Term 2:2-5.

CARTER, E.S. (1979) Comparison of Different Shrinkage Formulas in Estimating Population Multiple-Correlation Coefficients. Educational and Psychological Measurement, 39(2):261-6.

CHOPPIN, B.H.L.: ORR, L.: KURLE, S.D.M.: FARA, P; and JAMES, G. (1973) The Prediction of Academic Success. Sussex, Bucks; NFER Publishing Co. Ltd.

COSTIN, F. (1970) The Optimal Number of Alternatives in Multiple-Choice Achievement Tests: Some Empirical Evidence for a Mathematical Proof. Educational and Psychological Measurement, 30(2):353-8.

CUNNINGHAM, J.W. and TIERNEY, R.J. (1979) Evaluating Cloze as a Measure of Learning from Reading. Journal of Reading Behaviour, 11(3):287-92.

DAVIS, F.B. (1944) Fundamental Factors of Comprehension in Reading. Psychometrika, 9(3):185-97.

DAVIS, F.B. (1968) Research in Comprehension in Reading. Reading Research Quarterly, 3(4):499-545.

DEPARTMENT OF EDUCATION (1971) Report on the 1969 School Certificate Science Examination. Bulletin No. 51, Wellington; New Zealand Department of Education, Curriculum Development Unit.

DEPARTMENT OF EDUCATION (1984) Education Statistics of New Zealand. Wellington.

DUNCAN, R.E. (1983) An Appropriate Number of Multiple-Choice Item Alternatives: A Difference of Opinion. Measurement and Evaluation in Guidance, 15(4):283-92.

DUNN, S.S. (1977) Assessment for Tertiary Entrance. Education News, 15:12-4.

DUPUIS, M.M. (1980) The Cloze Procedure as a Predictor of Comprehension in Literature. Journal of Educational Research, 74(1):27-33.

DUPUY, H.P. (1974) The Rationale, Development and Standardization of a Basic Word Vocabulary Test. Washington, D.C.; U.S. Government Printing Office.

EBEL, R.L. (1972) Some Limitations of Criterion-Referenced Measurement. IN Bracht, G.H., Hopkins, K.D. and Stanley, J.C. (Eds.) Perspectives in Educational and Psychological Measurement. Englewood Cliffs, N.J.; Prentice-Hall:144-9.

EDUCATIONAL DEVELOPMENT CONFERENCE (1974) Assessment in Schools. A report prepared for the Working Party on Improving Learning and Teaching, Wellington.

EDUCATIONAL DEVELOPMENT CONFERENCE (1974) Directions for Educational Development. A report prepared by the Advisory Council on Educational Planning, Wellington.

EDUCATIONAL TESTING SERVICE (1972) Multiple Choice. IN Bracht, G.H., Hopkins, K.D. and Stanley, J.C. (Eds.) Perspectives in Educational and Psychological Measurement. Englewood Cliffs, N.J.; Prentice-Hall:150-6.

ELLEY, W.B. (1967) Estimating the Difficulty Level of Reading Material. Education, October:3-11.

ELLEY, W.B. (1969) The Assessment of Readability by Noun Frequency Counts. Reading Research Quarterly, 4(3):411-27.

ELLEY, W.B. (1976) Overseas Experience with Internal Assessment. Post Primary Teachers' Association Journal, April:26-32.

ELLEY, W.B. (1976a) The Cloze Procedure: A Method of Testing, Teaching and Assessing Readability. IN Doake, D.B. and O'Rourke, B.T. (Eds.) New Directions for Reading Teaching. Wellington; New Zealand Educational Institute.

ELLEY, W.B. (1977) A Close Look at the Cloze Test. Set, No. 1, Item 2. Wellington; New Zealand Council for Educational Research.

ELLEY, W.B. (1984) Exploring the Reading Difficulties of Second-Language Learners in Fiji. IN Alderson, J.C. and Urquhart, A.H. Reading in a Foreign Language. N.Y.; Longman: 281-97

ELLEY, W.B. (1985) Internal Assessment: The Issues. Post Primary Teachers' Association Journal, Term 1:8-15.

ELLEY, W.B. (In progress) Final Research Report on the National Survey of Teacher Opinion about the use of Reference Tests in Moderating Internal Assessment for School Leaving Awards. Christchurch; Education Department, University of Canterbury.

ELLEY, W.B. and LIVINGSTONE (1972) External Examinations and Internal Assessments. Wellington; New Zealand Council for Educational Research.

ELLEY, W.B. and LIVINGSTONE (Undated) A Proposal for Phasing Out School Certificate Examinations. Wellington; New Zealand Council for Educational Research.

ELLEY, W.B. and REID (1969) Progressive Achievement Tests in Reading Comprehension and Reading Vocabulary: Teacher's Manual. Wellington; New Zealand Council for Educational Research.

ELLEY, W.B., BARHAM, I.H., LAMB, H. and WYLLIE, M. (1979) Reliability of Essay Marking. IN The Role of Grammar in a Secondary School Curriculum. Wellington; New Zealand Council for Educational Research:88-95

ELWOOD, M.I. (1939) A Preliminary Note on the vocabulary Test in the Revised Stanford-Binet Scale. Journal of Educational Psychology, 30(8):632-4.

ENTIN, E. and KLARE, G. (1978) Some Interrelationships of Reliability, Cloze and Multiple-Choice Scores on a reading Comprehension Test. Journal of Reading Behaviour, 10(4):417-36.

FAIRBURN, K., McBRYDE, B. and RIGBY, R. (1976) Internal Assessment in Queensland. New Zealand Journal of Educational Studies, 11(2):143-51

FORSYTH, R.A. and SPRATT, K.F. (1980) Measuring Problem Solving Ability in Mathematics with Multiple-Choice Items: The Effect of Item Format on Selected Item and Test Characteristics. Journal of Educational Measurement, 17(1):31-43.

FRANKEL, E. (1960) Effects of Growth, Practice and Coaching on Scholastic Aptitude Test Scores. Personnel and Guidance Journal, 38:713-9.

FRARY, R.B. (1985) Multiple-Choice Versus Free-Response: A Simulation Study. Journal of Educational Measurement, 22(1):21-31.

GILMORE, A.M. (1979) The Use of Multiple Matrix Sampling Techniques in the Moderation of School Assessments. Unpublished Ph. D. Thesis (Education). Dunedin; University of Otago.

GILMORE, A.M. (1984) An Exercise in Moderation: Predicting School Certificate Performance via Multiple Matrix Sampling. New Zealand Journal of Educational Studies, 19(1):67-75.

GRIER, J.B. (1975) The Number of Alternatives for Optimum Test Reliability. Journal of Educational Measurement, 12(2):109-13.

GUILFORD, J.P. and FRUCHTER, B. (1978) Fundamental Statistics in Psychology and Education (6th Ed.) Tokyo; McGraw-Hill Kogakusha, International Student Edition.

GULLIKSEN, H. (1950) Theory of Mental Tests. New York; John Wiley & Sons.

HALL, C., McMURRAY, S. and CAPPER, P. 91985) Sixth Form Certificate Grade Allocation. Post Primary Teachers' Association Journal, Term 1:34-7.

HARGIS, C.H. (1972) A Comparison of Retarded and Nonretarded Children on the Ability to Use Context in Reading. American Journal of Mental Deficiency, 76(6):726-8.

HEIM, A.M. and WATTS, K.P. (1967) An Experiment on Multiple-Choice Versus Open-Ended Answering in a Vocabulary Test. British Journal of Educational Psychology, 37(3):339-46.

HELFELDT, J.P., HENK, W.A. and FOROS, A. (1986) A Test of Alternative Cloze Test Formats at the Sixth Grade Level. Journal of Educational Research, 79(4):216-21.

HENK, W.A. (1981) Effects of Modified Deletion Strategies and Scoring Procedures on Cloze Test Performances. Journal of Reading Behaviour, 13(4):347-57.

HENRYSSON, S. (1964) The Swedish System for Equalizing Marks. Educational Research, 6:156-60.

HOGAN, H.M. (1976) Replacing School Certificate. Post Primary Teachers' Association Journal, April:22-5.

HOSSEINI, J. and FERRELL, W.R. (1982) Measuring Metacognition in Reading by Detectability of Cloze Accuracy. Journal of Reading Behaviour, 14(3):263-74

HUDSON, B. (1973) (Ed.) Assessment Techniques: An Introduction. London; Methuen Educational.

HUGHES, D.C. and KEELING, B. (1976a) (1976b) Internal Assessment for School Certificate. Post Primary Teachers' Association Journal, Part 1, February: 42-4; Part 2, March:38-9.

HULBERT, M.E. (1978) Moderating Internal Assessment for School Certificate. Unpublished M. Phil. Thesis, University of Waikato.

JENKINSON, M.D. (1957) Selected Processes and Difficulties of Reading Comprehension. Unpublished Ph. D. Thesis, University of Chicago.

KEEPES, B.D. and KEEPES, J.M. (1974) Practices in the USA and their Relevance to Australia. IN Dunn, S.S. (Ed.) Public Examinations: The Changing Scene. Adelaide; Rigby: 181-205.

KEEVES, J.P., McBRYDE, B. and BENNETT, L.A. (1977) The Validity of Alternative Methods of Scaling School Assessments of the Australian Capital Territory. Paper presented at the Annual Conference of the Australian Association for Research in Education, Canberra.

KELLY, T.R. (1927) Interpretation of Education Measurements. Yonkers-on-Hudson, N.Y.; World Book.

LEES, L. (1979) Research Relating to the Australian Scholastic Aptitude Test: A Selected Annotated Bibliography. Hawthorn, Victoria; The Australian Council for Educational Research.

LEVINE, A.S. (1958) Aptitude Versus Achievement Tests as Predictors of Achievement. Educational and Psychological Measurement,18(3):517-25.

LEWINSKI, R.J. (1948) Vocabulary and Mental Measurement: A Quantitative Investigation and Review of Research. Journal of Genetic Psychology, 72:247-81.

LORD, F.M. (1962) Estimating Norms by Item Sampling. Educational and Psychological Measurement, 22:259-67.

MacINTOSH, H.G. and MORRISON, R.B. (1969) Objective Testing. London; University of London Press.

MacKAY, L.D. and FARY, B. (1979) Testing in Transition. Report on the Tertiary Education Entrance Project in Victoria, S.A.

MARANDOS, S.A. (1974) Analysis of the Cloze Procedure as a Measure of Reading Comprehension. Unpublished M.A. Thesis, California State University.

MARKLAND, S. (1985) Education in Sweden: Assessment of Student Achievements and Selection for Higher Education. University of Stockholm, Sweden.

MATTHEWS, D.A. (1983) The Use of Standardized Tests: A Practitioners Point of View. New Zealand Journal of Educational Studies, 18(2):171-8.

McCAUSLAND, F.J. (1981) Teachers' Predictions and Pupil Achievement in School Certificate. Unpublished M.A. Thesis (Education), Victoria University of Wellington.

McCAUSLAND, F.J. and HALL, C.G.W. (1985) The Accuracy of Teachers' Predictions of Pupils' Marks in School Certificate. New Zealand Journal of Educational Studies, 20(1):82-92.

McCLELLAND, W. (1949) Selection for Secondary Education. London; University of London Press.

McCOMBS REPORT (1976) Towards Partnership. Report on the Committee on Secondary Education. Wellington; Department of Education.

McGAW, B. (1974) Internal Assessment and the Problems of Moderation. Post Primary Teachers' Association Journal, May:33-7.

McGAW, B. (1976) The Napier Internal Assessment Scheme. Post Primary Teachers' Association Journal, November:21-3.

McGAW, B. (1977) ,The Use of Rescaled Teacher Assessments in the Admission of Students to Tertiary Study. The Australian Journal of Education, 21:209-25.

McGAW, B. WARRY, R. and McBRYDE, B. (1975) Validation of Aptitude Measures for the Rescaling of School Assessments. Education Research and Perspectives, 2(2):20-34.

McGAW REPORT (1984) Assessment in the Upper Secondary School in Western Australia. Ministerial Working Party on School Certification and Tertiary Admissions Procedures, Perth.

McGAWE, G.M. (1974) The Effects of Test-Wiseness on Achievement in Multiple-Choice and Short Supply Item Tests in Fifth Form Science. Unpublished M.A. Research Paper (Education). Christchurch; University of Canterbury.

McKENNA, M. (1976) Synonymic Versus Verbatim Scoring of the Cloze Procedure. Journal of Reading, 20(2):141-3.

MURPHY, R.J.L. (1979) Teachers' Assessments and GCE Results Compared. Educational Research, 22:54-9.

MURPHY, R.J.L. (1981) O-Level Grades and Teachers' Estimates as the Predictors of the A-Level Results of UCCA Applicants. British Journal of Educational Psychology, 51:1-9.

NITKO, A.J. (1983) Educational Tests and Measurement - An Introduction. New York; Harcourt Brace Jovanovich.

NUTTALL, D.L. (1971) The 1968 CSE Monitoring Experiment. Schools Council Working Paper 34. London; Evans Brothers.

OTTO, E.P. (1976) ASAT and Other Factors Related to Academic Performance. Education Research and Perspectives, 3(1):34-44.

PETCH, J.A. (1964) School Estimates and Examination Results Compared. Manchester; JMB.

PLUMLEE, L.B. (1964) Estimating Means and Standard Deviations from Partial Data - An Empirical Check on Lord's Sampling Technique. Educational and Psychological Measurement, 24(3):623-30.

POWER, C. (1986) Criterion-Based Assessment, Grading and Reporting at Year 12 Level. Australian Journal of Education, 30(3):266-84.

PUBLIC EXAMINATIONS BOARD OF SOUTH AUSTRALIA (1975) Report of ASAT Committee, November.

QUEENSLAND BOARD OF SECONDARY SCHOOL STUDIES (1978) Analysis of ASAT SERIES F as used in Queensland, 1977. IN Research Papers Relating to the Australian Scholastic Aptitude Test. Hawthorne, Victoria; Australian Council for Educational Research: 43-8.

QUEENSLAND BOARD OF SECONDARY SCHOOL STUDIES (1978a) A Study of correlation between ASAT scores and school assessments in a sample of schools in Queensland 1975-77. IN Research Papers Relating to the Australian Scholastic Aptitude Test. Hawthorne, Victoria; Australian Council of Educational Research: 49-60.

RADFORD, W.C. (1974) Trends in Australia, New Zealand and Scotland. IN Dunn, S.S. (Ed.) Public Examinations: The Changing Scene. Adelaide; Rigby.

RAMOS, R.A. and STERN, J. (1973) Item Behaviour Associated with the Changes in the Number of Alternatives in Multiple-Choice Items. Journal of Educational Measurement, 10(4):305-10.

RANKIN, E.F. and CULHANE, J.W. (1969) Comparable Cloze and Multiple-Choice Comprehension Test Scores. Journal of Reading, 13(3):193-8.

RANKIN, E.F. and DALE, L.H. (1969) Cloze Residual Gain - A Technique for Measuring Learning Through Reading. IN Schick, G.B. and May, M.M. (Eds.) The Psychology of Reading Behaviour. Milwaukee, Wisconsin; 18th Yearbook of the National Reading Conference: 17-26.

RATNAMALAR, S. (1986) Assessing the Difficulty Level of Textbooks Used by Secondary One Pupils in Singapore Using the Cloze Procedure. Unpublished M. Ed. Research Paper (Education). Chrsitchurch; University of Canterbury.

REID, N.A. and HUGHES, D.C. (1974) Factor Analysis of the PAT Reading and Listening Tests. New Zealand Journal of Educational Studies, 9(1):18-30.

REID, N.A., JACKSON, P., GILMORE, A.M. and CROFT, C. (1981) Test of Scholastic Abilities: Teacher's Manual. Wellington; New Zealand Council for Educational Research.

ROSENBERG, J. (1976) The Use of ASAT for Rescaling School Assessments. Educational Research and Perspectives, 3(1):26-33.

ROSS REPORT (1986) Learning and Achieving: Secondary Report of the Committee of Inquiry into Curriculum, Assessment and Qualifications in Forms 5 to 7. Wellington; Department of Education.

ROWLANDS, R.G. (1974) Moderation. IN Dunn, S.S. (Ed.) Public Examinations: The Changing Scene. Adelaide; Rigby:85-110.

ROWLEY, G.L. (1974) Which Examinees are Most Favoured by the Use of Multiple-Choice Tests? Journal of Educational Measurement, 11(1):15-23.

SAX, G. (1980) Principles of Educational and Psychological Measurement (2nd Ed.) Belmont, California; Wadsworth.

SCHOOL CERTIFICATE EXAMINATION BOARD (1974) Internal Assessment for School Certificate. Wellington; Department of Education.

SCHOOLS COUNCIL (1965) The Certificate of Secondary Education: School-Based Examinations: Examining, Assessing and Moderating by Teachers. Examinations Bulletin No. 5. London; HMSO.

SCHOOLS COUNCIL (1966) The 1965 CSE Monitoring Experiment, Parts I and II. Working Paper 6. London; HMSO.

SCHOOLS COUNCIL (1972) The Predictive Value of CSE Grades for Further Education. Examinations Bulletin No. 24. London; Evans/Methuen Educational.

SHOEMAKER, D.M. (1970a) Item-Examinee Sampling Procedures and Association of Standard Errors in Estimating Test Parameters. Journal of Educational Measurement, 7:255-62.

SHOEMAKER, D.M. (1970b) Allocation of Items and Examinees in Estimating a Norm Distribution by Item Sampling. Journal of Educational Measurement, 7:123-8.

SHOEMAKER, D.M. (1973) Principles and Procedures of Multiple Matrix Sampling. Cambridge, Mass; Ballinger Publishing.

SKURNIK, L.S. and HALL, J. (1968) The 1966 CSE Monitoring Experiments. Schools Council Working Paper 21. London; HMSO.

SLANE, J.V.G. (1968) An Examination of Some Readability Measures. Unpublished M.A. Thesis (Education). Auckland; University of Auckland.

SMITH, F. (1971) Understanding Reading. New York; Holt, Rhinehart and Winston.

SOWELL, D.J.W. (1970) CSE Grades and Teacher Forecasts. Educational Research, 13:28-35.

SPSSX (1983) Statistical Packages for the Social Sciences (User's Guide). Chicago, Illinois; SPSS Inc.

STRATON, R.G. and GATTS, R.M. (1980) A Comparison of Two, Three And Four-Choice Item Tests Given a Fixed Total Number of Choices. Educational and Psychological Measurement, 40(2):357-65.

SUTHERLAND, J.E.N. (1972) The Tertiary Education Entrance Project. Education News, 13(9):23-8.

SWANSON, R.G. (1976) Multiple Choice Tests: How Many Alternatives? Maxwell AFB, Ala; Academic Instructor School.

TAYLOR, W.L. (1953) Cloze Procedure: A New Tool for Measuring Readability. Journalism Quarterly, 30:415-33.

TERMAN, L.M. (1918) Vocabulary Test as a Measure of Intelligence. Journal of Educational Psychology, 9:452-466.

THOMSON, J.D. and SLEE, C.W. (1975) The Predictive Validity of the Commonwealth Secondary Scholarship Examination Tests. CSSE Research Report No. 5. Hawthron, Victoria; Australian Council of Educational Research.

THONDIKE, R.C. (1971) (Ed.) Educational Measurement (2nd Ed). Washington, D.C.; American Council of Education.

TRAUB, R.E. and FISHER, C.W. (1977) On the Equivalence of Multiple-Choice and Free-Response Tests. Applied Psychological Measurement, 1(3):355-69.

TVERSKY, A. (1964) On the Optimal Number of Alternatives of a Choice Point. Journal of Mathematical Psychology, 1(2):386-91.

WARD, W.C. (1982) A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests. Applied Psychological Measurement, 6(1):1-11.

WILLMONT, A.S. and HALL, C.G.W (1975). O-Level Examined: The Effect of Question Choice. London; McMillan Education.

WILSON, J.M. (1982) The Accuracy of A-Level Forecasts. Educational Research, 24:216-22.

YATES, A. (1953) Symposium of the Effects of Coaching and Practice In Intelligence Tests. British Journal of Educational Psychology, 23(3):147-62.

YATES, A. and PIDGEON, D.A. (1957) Admission to Grammar Schools. NFER Publication No. 10, Newnes Educational.